OPTIMAL USE OF PHENOTYPIC DATA FOR BREEDING USING GENOMIC PREDICTIONS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Nicolas Didier Heslot

January 2014

OPTIMAL USE OF PHENOTYPIC DATA FOR BREEDING USING GENOMIC PREDICTIONS

Nicolas Didier Heslot, Ph. D.

Cornell University 2014

Genomic predictions or genomic selection (GS) was proposed to overcome a number of challenges in application of marker assisted selection to complex quantitative traits. Simulations and empirical studies suggest that GS can improve genetic gain per unit time and cost. The cost of molecular markers has dramatically decreased over the past 10 years and should continue to do so with progress in sequencing technologies whereas phenotyping cost should remain stable or increase with land and labor costs. This means that the most valuable and limiting part in breeding will increasingly be the phenotype and not the genomic data. As a consequence, it is critical to make the most of the scarce phenotypic data available. GS opens numerous possibilities to do so.

First, using eight wheat (*Triticum aestivum* L.), barley (*Hordeum vulgare* L.), *Arabidopsis thaliana* (L.) Heynh., and maize (*Zea mays* L.) datasets, the predictive ability of currently available GS models was evaluated by comparing accuracies, the genomic estimated breeding values (GEBVs), and the marker effects for each model. While a similar level of accuracy was observed for many models, the computation time varied widely as did the distribution of marker effect estimates.

Second, allele replication rather than genotype replication was investigated as a new way to cope with highly unbalanced phenotypic data sets. Using a two-row elite barley

(*Hordeum vulgare* L.) population from a commercial breeding program, I demonstrated the possibilities offered by GS to analyze multienvironment trials, identify outliers, group environments, and select historical data relevant for current breeding efforts.

Finally, we proposed, developed and tested a new model to use environment data to model genotype by environment interactions (G*E) in GS. A crop model was used to derive stress covariates from daily weather data for predicted crop development stages. I extended the factorial regression model to genomic selection. Machine learning was also used to capture non-linear responses of QTL to stresses. The method was tested using a large winter wheat dataset. This new model provides insight into the genetic architecture of genotype by environment interactions and could predict genotype performance based on past and future weather scenarios.

BIOGRAPHICAL SKETCH

Nicolas Heslot was born in Chicago, IL in 1986 but grew up in Viroflay, France. From an early age he had a keen interest in understanding how plants grew. His parents claim that at the age of 3 he was uprooting plants in the backyard to see how they were made. He graduated from Saint Jean Hulst high school in Versailles, France, majoring in mathematics in 2004. He then studied for two years at Lycée Hoche in Preparatory classes for the national selective examination for admission to the French "Grandes Ecoles" (graduate-level engineering schools) majoring in Mathematics and Biology. Those two years were heavily dedicated to fundamental, biology, physics, chemistry and mathematics classes. At the national entrance examination in 2006, he ranked 17th out of 2624 and was admitted at AgroParisTech (formerly INA P-G), Paris Institute of Technology for life, food and environmental sciences, France's top degree program, engineering school for agricultural and life sciences. There he started learning more about agriculture, agronomy and engineering. As part of the curriculum he thoroughly enjoyed working full time for two months in a dairy farm near Le Mans (France). It is also at AgroParisTech that he took introduction courses in quantitative and population genetics. He really appreciated the system approach and the use of mathematics to inform biological research. This convinced him to work in the seed industry. However, he was also interested in business and took a year off from school in 2008/2009 to decide which direction to pursue. He worked for 7 months as a marketing assistant for Limagrain Europe (then Limagrain Verneuil Holding) in Verneuil l'Etang, France. Under the direction of Laurent Wilsdorf, European marketing and development manager for Limagrain Europe he studied the non-food use of sunflower and rapeseed, in particular for biodiesel. He investigated if those uses required specific seed qualities

that could be delivered by new varieties, the potential market size and if industrial users would pay a premium for it.

He then moved to Syngenta Biotechnology, (Research Triangle Park, NC) where he worked with Mary-Dell Chilton (World Food Prize 2013 recipient) on improving *agrobacterium* mediated maize transformation. He focused on creating alternative helper plasmids with new genes to try to suppress the cell death induced by *agrobacterium* in some maize elite inbreds. He loved this research experience and thanks Mary-Dell Chilton for convincing him to pursue a research career. However, he wanted to have a broader picture approach than with molecular biology and went back to school to finish a Master of Science and engineering degree majoring in plant genetics and breeding in 2010. He performed the research for his Master thesis at CIMMYT in Mexico under the supervision of David Bonnett, working on novel sources of resistance to fusarium head blight and *Septoria tritici* in synthetic breadwheat. When he was in Mexico, Limagrain Europe knowing that he was looking to pursue a PhD offered to fund his graduate studies at Cornell. He joined Cornell University to pursue a PhD in plant breeding and genetics in January 2011. Upon graduation he is going back to Limagrain Europe to work on marker-assisted breeding innovation.

For Mathilde

« Ab eo qui fecit te noli deficere nec ad te »

« De celui qui t'a fait, ne t'éloigne pas, même pour aller vers toi»

Saint Augustin, De Continentia

« I see our scientific theories as human inventions – nets designed by us to catch the world. To be sure, these differ from the inventions of the poets, and even from the inventions of the technicians. Theories are not only instruments. What we aim at is truth: we test our theories in the hope of eliminating those which are not true. In this way we may succeed in improving our theories – even as instruments: in making nets which are better and better adapted to cach our fish, the real world. »

Karl Popper, The open universe, in Postscript to the logic of scientific discovery, 1956

ACKNOWLEDGMENTS

I would like to thank my advisors **Dr. Mark Sorrells** and **Dr. Jean-Luc Jannink** for giving me the freedom to pursue my research interests. I also thank them for their continuous support and high expectations. I have greatly benefited from their wisdom, guidance, patience and friendship.

I would also like to acknowledge **Dr. Jason Mezey,** my other thesis committee member for stimulating and challenging feedback.

Special thanks are due to **Dr. Mary-Dell Chilton** for convincing me to pursue scientific endeavors.

Cornell University is a great place to conduct research. The faculty, staff and students in the **Department of Plant Breeding and Genetics** were essential to making my Ph.D. experience an amazing one.

Thanks are also due to past and present members of the Jannink and Sorrells labs for for their support and friendship. In particular I would like to acknowledge the great support of **Dr. Jeffrey Endelmann.**

I am extremely grateful for the continuing support of **Limagrain Europe** through funding, access to data and many inspiring discussions: In particular, **Laurent Wilsdorf** who first hired me in the company, **Stéphanie Chauvet** who was always available to help me with computation and R issues. **Anne-Marie Bochard**, **Jayne Stragliati** and **Bruno Poupard** provided data for my research and gave me great encouragements and feedback. Finally, I wish to express my deepest gratitude to **Pascal Flament** for giving me the opportunity to pursue a Ph.D. at Cornell.

I would also like to thank all of my supportive friends and colleagues. Most importantly, I would like to thank my family for their support.

TABLE OF CONTENTS

BIOGRAPHICAL SKETCH iii
ACKNOWLEDGMENTS vi
TABLE OF CONTENT vii
LIST OF FIGURES viii
LIST OF TABLES x

CHAPTER 1 PERSPECTIVES FOR GENOMIC SELECTION APPLICATIONS AND RESEARCH IN PLANTS
ABSTRACT 1
REFERENCES 23

CHAPTER 2 GENOMIC SELECTION IN PLANT BREEDING: A COMPARISON OF MODELS
ABSTRACT 30
MATERIALS AND METHODS 33
RESULTS 49
DISCUSSION 64
REFERENCES 71

CHAPTER 3 USING GENOMIC PREDICTION TO CHARACTERIZE ENVIRONMENTS AND OPTIMIZE PREDICTION ACCURACY IN APPLIED BREEDING DATA
ABSTRACT 76
MATERIALS AND METHODS 79
RESULTS 89
DISCUSSION 96
REFERENCES 103

CHAPTER 4 IMPACT OF MARKER ASCERTAINMENT BIAS ON GENOMIC SELECTION ACCURACY AND ESTIMATES OF GENETIC DIVERSITY
ABSTRACT 107
MATERIALS AND METHODS 110
RESULTS 117
DISCUSSION 126
REFERENCES 130

CHAPTER 5 INTEGRATING ENVIRONMENTAL COVARIATES AND CROP MODELING INTO THE GENOMIC SELECTION FRAMEWORK TO PREDICT GENOTYPE BY ENVIRONMENT INTERACTIONS
ABSTRACT 133
MATERIALS AND METHODS 140
RESULTS 160
DISCUSSION 167
REFERENCES 175

LIST OF FIGURES

Figure 1.1. Key parameters and changes during a breeding cycle, to consider in implementing GS 6

Figure 1.2. Simple scheme of a breeding program with for each stage what GS could bring in orange. 8

Figure 1.3. Different sources of information for performance prediction and their integration in a breeding program. 16

Figure 2.1. A generic feed-forward neural network with a single hidden layer 41

Figure 2.2. Heat maps summarizing accuracies for the grid search on weighted Bayesian shrinkage regression (wBSR) prior parameters 51

Figure 2.3. Heat map summary of the grid search on empirical Bayes (E-Bayes) prior parameters 53

Figure 2.4. Hierarchical clustering of genomic selection (GS) models based on cross-validated genomic estimated breeding values (GEBVs) 57

Figure 2.5. Comparison of marker effect distribution. 59

Figure 3.1. Representation of the optimization procedure used 87

Figure 3.2. Relationship between the mean prediction accuracy of the environments using one environment for training and the training population size 90

Figure 3.3. Heat map of environments based on Euclidean distances computed using marker effects. 91

Figure 3.4. Heat map of the prediction accuracy between pairs of environments excluding common lines. 92

Figure 3.5. Results of the training population optimization approach 95

Figure 4.1. PCA plots for respectively all DArT markers (A) and all GBS markers available (B) 118

Figure 4.2. Heatmap of the R^2 of the eigenvector between the two platforms 119

Figure 4.3. DArT MAF distribution and 95% confidence interval from the GBS bootstrap 122

Figure 4.4. DArT PCA R^2 and 95% confidence interval from the GBS bootstrap 123

Figure 5.1. Typical structure of a crop model 137

Figure 5.2. Major stresses by development stage for winter wheat. 139

Figure 5.3. Modeling flow diagram 149

Figure 5.4. A simple regression tree built on a bootstrap sample of observations and variables. 154

Figure 5.5. Distribution of the variance of marker effects, computed in each environment, across environments, plotted on a log scale. 161

Figure 5.6. Predictive performances of the models as a function of the number of markers for which a linear sensitivity to each covariate is fitted from the cross-validation. 162

Figure 5.7. Comparison of the prediction accuracy in each predicted environment from the cross validation between the main effect model and the model with rules of order four and 250 markers for the factorial regression added to the main effect prediction. 163

Figure 5.8. Hierarchical agglomerative clustering of the environments based on the predicted G*E response of all genotypes. 166

Figure 5.9. Heatmap of the correlation of G*E predicted values between environments after hierarchical agglomerative clustering of the environments 173

LIST OF TABLES

Table 2.1. Dataset origins and details. 34

Table 2.2. Accuracy for each trait and model, average non-cross-validated correlation for each model, and average MSE for each model. 55

Table 2.3. Summary of the subpopulation results 61

Table 3.1. *P*-value of Mantel tests between the environments' Euclidean distance matrices based on marker effects, the accuracy matrices between environments, and different measures of genetic distance between environments 93

Table 4.1. Population genetics parameters computed using the DArT and p-value from the GBS bootstrap. 120

Table 4.2. Non-redundant GBS and DArT markers and P-value function of the R^2 cutoff. 124

Table 4.3. Cross-validated GS accuracy for DArT and GBS and boostrap p-values for the DArT markers. 125

Table 5.1. Stress covariates used and references, modified from Lecomte (2005) 143

Table 5.2. Importance of the eight stress covariates with a significant importance, for the model with rules of order 4 and 250 markers for the factorial regression. 165

CHAPTER 1

PERSPECTIVES FOR GENOMIC SELECTION APPLICATIONS AND RESEARCH IN PLANTS

Abstract

Genomic selection (GS) has created a lot of excitement and expectations in the animal and plant breeding research communities. In this review, we briefly describe how genomic prediction can be integrated into breeding efforts and point out achievements and areas where more research is needed. GS provides many opportunities to increase genetic gain in plant breeding per unit time and cost. Early empirical and simulation results are promising, but for GS to deliver genetic gains, a systems perspective, as well as, careful consideration of the problem of optimal resource allocation is needed. This means considering the cost-benefit balance of using markers for each trait and stage of the breeding cycle instead of only focusing on recurrent selection with GS on a few complex traits using prediction on unphenotyped individuals. With decreasing marker cost, phenotype data is quickly becoming the most valuable asset and marker-assisted selection strategies should focus on making the most of scarce and expensive phenotypes. It is important to realize that markers can also improve accuracy of selection for phenotyped individuals. Use of markers as an aid to phenotype analysis suggests a number of new strategies in terms of experimental design and multi-trait models. GS also provides new ways to analyze and deal with genotype by environment interactions. Lastly, I point to some recent results showing that new models are needed to improve predictions particularly with respect to the use of distantly related individuals in the training population.

Introduction

Use of molecular markers as an aid to selection has been an active area of research for several decades now (Lande and Thompson 1990; Stuber et al. 1982; Tanksley et al. 1989) and generated a lot of expectations but early results have been disappointing for complex quantitative traits (Moreau et al. 2004). The concept of genomic selection (GS) by (Meuwissen et al. 2001) fostered great hopes and opened new ways to use molecular markers in breeding for complex traits. Initially, most of the research was conducted in the animal breeding community, where the high cost of phenotyping (e.g. progeny testing in dairy cattle breeding), as well as the impossibility to replicate individuals, made it attractive. In addition, because of that impossibility to replicate individuals, animal breeders implemented mixed model methodology early on to analyze their data using the available pedigree information (Henderson 1984). In plant breeding, the use of mixed models is more recent and not yet as widespread (Piepho et al. 2007; Smith et al. 2005). As a consequence, organizations, infrastructures and people were more prepared to embrace GS in animal breeding than in the plant community. I briefly summarize what is known about GS in plants, advocate for a systematic approach in the use of markers, and attempt to identify where GS could deliver improved genetic gains beyond recurrent selection. I also point out areas where more research is needed for GS to effectively deliver increased genetic gain per unit time and cost.

What is known

Marker-assisted recurrent selection (MARS) is a large class of breeding schemes using markers to select unphenotyped individuals and quickly cross them to generate another generation of candidates. Initial work with MARS used biparental or multi-populations QTL detection and then tried to pyramid them (Servin et al. 2004).

Some genomic selection reports make a distinction between MARS and GS but it is more logical to consider GS as a tool to carry out MARS among other possible uses.
I define genomic selection or prediction as the simultaneous use of genome-wide markers to predict an individuals' own performance or breeding value. This applies to both observed and unobserved individuals. GS can also make use of information on correlated traits to improve prediction accuracy by the use of multi-trait models. Early work on GS in plants was mainly focused on unobserved individuals, in the MARS context. But it can be beneficial for observed individuals as well if heritability is low (Endelman et al. 2013). The success of a breeding program is based on the release of improved cultivars and the way they were obtained does not matter. As a consequence, for a successful breeding program, resource allocation must be considered overall.

GS can be performed with a variety of statistical methods as reviewed in (Lorenz et al. 2011). Those methods are concerned with the same so called "large p small n" problem: There are many more predictor (marker) effects to be estimated than observations. Most of those approaches involve some type of penalized regression. Early research on GS in plants focused on prediction power measured through cross-validated accuracy using existing data (Heffner et al. 2011; Heslot et al. 2012; Lorenzana and Bernardo 2009) and on ensuring that GS was potentially more effective than classical marker-assisted selection schemes or use of pedigree (Asoro et al. 2013; Crossa et al. 2010). Results clearly indicated that GS was more predictive than classical marker-assisted selection in cross-validation and that with empirical data all GS methods had very similar prediction power. Currently, the most widely used model is the genomic best linear unbiased prediction model (GBLUP) (Habier et al. 2007). With GBLUP, markers are used to estimate the covariance between individuals. That information is further used in a mixed model analysis to predict

performance of observed and unobserved individuals. The GBLUP model has the advantage of relative simplicity, limited computing time and well-known optimality properties.

GS can greatly shorten the selection cycle (Heffner et al. 2010; König et al. 2009; Schaeffer 2006) and thus increase genetic gain per unit time and cost compared to phenotypic selection. A shortened selection cycle raised concerns that GS might increase the rate of loss of genetic diversity (inbreeding) and negatively impact long term selection gain. In simulation of recurrent selection with GS, without model updating, (Jannink 2010) showed that long term gain is reduced compared to phenotypic selection because GS cannot take into account rare alleles. Appropriate weighting of rare alleles can be used to preserve long term gains (Goddard 2009). Those studies focused on recurrent selection without model updating. In practice, because of the steep decline in accuracy with cycles of selection (Long et al. 2011), the model will probably be frequently updated with new phenotypic information. This should limit the problem of long term gains and inbreeding but warrant further investigations with simulations.

Currently, most of the selection in plants is based on phenotypic data collected on the selection candidates themselves with little or no use of pedigree. GS is effectively using information on relatives, through markers to carry out selection. As a consequence, the rate of inbreeding should increase, even if the GS models capture part of the Mendelian sampling. The Mendelian sampling is the genetic difference between individuals with the same pedigree, such as full-sibs. However, in animals, because previously the selection used pedigrees, which do not capture the Mendelian sampling, the use of GS should decrease the rate of inbreeding.

Research has also focused on training population size, the marker type and density required for GS (reviewed in (Lorenz et al. 2011). Briefly, larger training populations and higher marker density are beneficial in theory. In practice, accuracy usually reaches a plateau with increased marker number (Lorenz et al. 2012) and larger training populations do not always generate higher prediction accuracies (Riedelsheimer et al. 2013). Marker type and potential ascertainment bias have a limited impact on prediction accuracy as long as markers are at high density and well distributed across the genome (Heslot et al. 2013c). Finally, (Heffner et al. 2009; Lorenz et al. 2011) pointed out that with GS, phenotyping is done to train a model, not to directly select. As a consequence, the unit of evaluation is not the individual but the allele. This has raised questions on how to best design training populations under budget limitations (Rincent et al. 2012). This also suggested new ways to deal with unbalanced historical data where alleles will be replicated across environments (Heslot et al. 2013b). In the following, I argue that the concept that the unit of evaluation is the allele rather than the individual is meaningful only for marker-assisted recurrent selection (MARS) with GS.

Urgent need for a systems approach

Reflecting upon the disappointing results of marker-assisted selection, the apparent lack of success was due not only to inadequate statistical methodology for complex traits, for which GS provides a solution, but also to a number of practical problems that GS has not eliminated (Bernardo 2008; Xu and Crouch 2008). Among those practical problems are the choice of germplasm to apply MAS, integration of information on multiple traits, trade-off between population sizes and number of populations created for MAS, balance between phenotypic selection and MAS at constant budget, disconnection between the population used to detect QTLs and the

elite breeding germplasm and logistical issues in the integration of MAS in breeding programs.

Behind most of those issues is the problem of resource allocation between phenotyping and genotyping. Figure 1.1 presents the key parameters to consider when implementing the use of markers in a plant breeding program. Most key variables are trait-specific and vary during the breeding cycle.

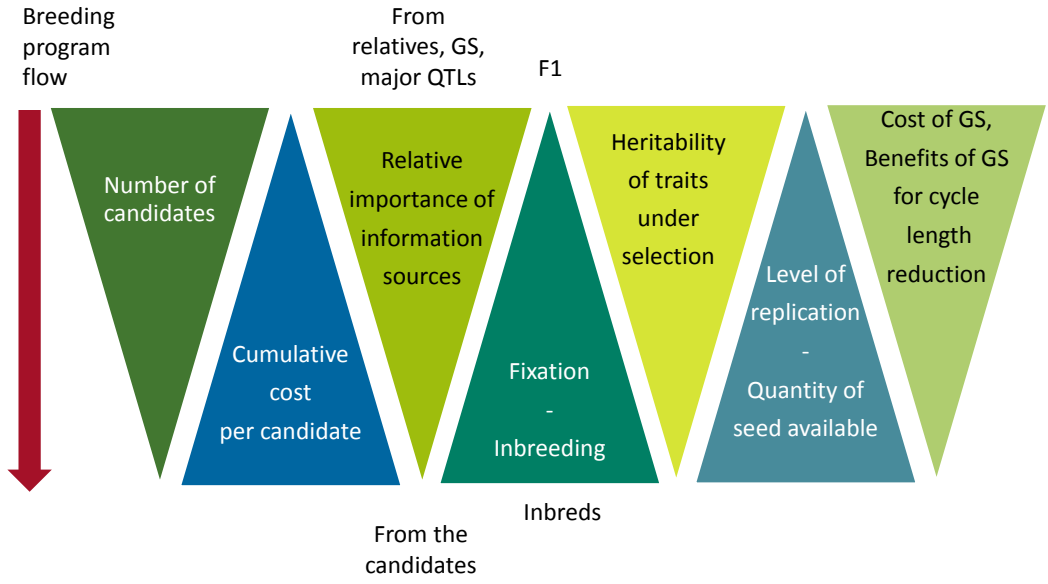


Figure 1.1. Key parameters and changes during a breeding cycle, to consider in implementing GS. The triangles indicate increase or decrease of the quantity considered.

This reveals a trade-off between an increased benefit of using markers to select early in the cycle on low heritability traits such as yield, and thereby reduce the length of the cycle, versus a higher cost of GS applied in early generations. The increased cost arises because the selection candidates are much more numerous and they are not fully inbred, making the logistics of genotyping and prediction more complicated. This trade-off is even stronger in a phenotypic breeding program, because large populations early in the cycle are combined with high selection intensity on highly heritable traits, which can be extremely efficient and relatively inexpensive. It is probably beneficial

to use markers to select on a low heritability trait such as yield early in the cycle. In most crops, yield cannot be measured accurately on segregating populations, single plants or small plots. At the same time, most of the individuals in early generations can be discarded efficiently using inexpensive phenotyping. An extreme example of the usefulness of at least limited phenotyping in early generations is dairy cattle. (Hayes et al. 2009b) pointed out that phenotype is still needed on all candidates before release to eliminate congenital defects caused by rare alleles becoming homozygous and novel mutations with large effects. These requirements should be taken into account in implementing GS in a breeding program.

Effective use of markers to achieve breeding gains requires a systems perspective. Implementing GS in dairy cattle can generate enough savings in phenotyping expenses to pay for the genotyping. (Schaeffer 2006) estimated the cost of testing a bull to be 50,000 dollars. In other animal species, cost effective GS requires complex optimization strategies. In pig, (Tribout et al. 2013) found that implementing GS was beneficial only if the breeding program budget was greatly increased. In their simulations, below a given threshold, additional resources were better allocated to more phenotyping. In salmon, (Lillehammer et al. 2013) investigated pre-selection of candidates based on pedigree before GS to limit costs. Similar studies are needed in plants.

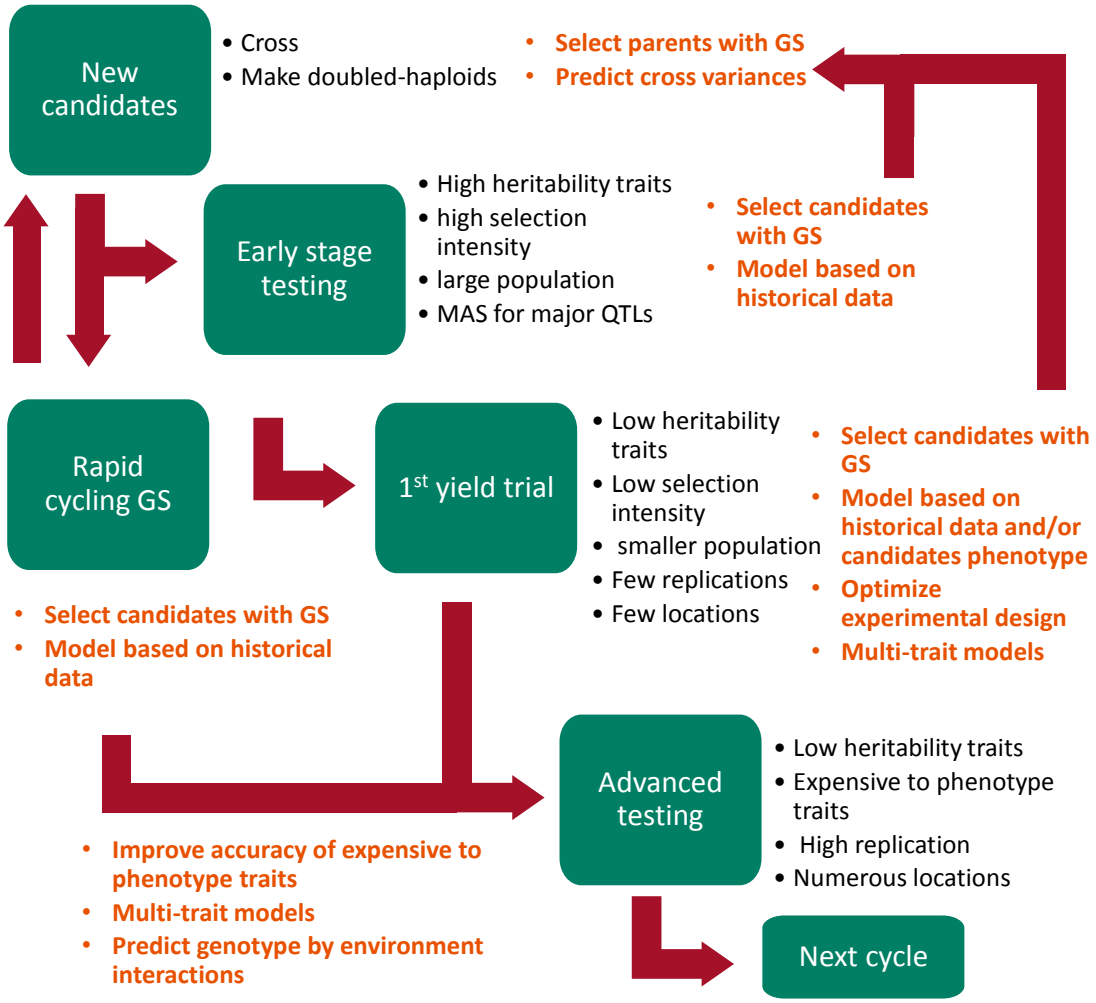


Figure 1.2. Simple scheme of a breeding cycle with what GS could bring for each stage (orange). Arrows indicate the flow of germplasm. Upward arrows correspond to early recrossing. For the sake of simplicity, the scheme uses doubled-haploids. MAS: marker-assisted selection.

Figure 1.2 presents a schematic of breeding inbred lines using doubled-haploids**.** For each stage, the figure presents side by side characteristics of classic breeding (in black) and potential applications of GS (in orange). Justification for specific GS applications and associated research needs are presented later in this review.

Clearly, GS can be of some use at each stage of cultivar development. Nevertheless, most of the attention and empirical validation of GS has been focused

only on marker-assisted recurrent selection (Rapid cycling GS on Figure 1.2) for one trait (Bernardo and Yu 2007; Moreau et al. 2004).

The relevance of those applications depends on budget size and relative costs of phenotyping and genotyping. A systems perspective is needed to leverage the strength of both phenotypic and marker-assisted selection. It could be achieved by simulations to compare multiple GS breeding strategies at constant budget for example (Endelman et al. 2013). A decisive step towards better resource allocation would also be the ability to identify the most promising crosses based on expected mean and variance (Zhong and Jannink 2007).

Marker-assisted recurrent selection with GS

Early molecular breeding efforts were based on quantitative trait loci (QTL) mapping in biparental or connected populations for a few traits of interest. As a consequence, practical application focused on recurrent selection schemes aided by markers to quickly pyramid identified QTLs (Servin et al. 2004). With the advent of GS, effort have been devoted to make those recurrent selection schemes work using GS (Bernardo 2009; Bernardo and Yu 2007). Usually, a narrow based population is created, such as a biparental population or connected crosses and phenotyped or predicted with a GS model built using historical data. Some individuals are selected and quickly intermated and used directly for more advanced testing (Rapid cycling GS on Figure 1.2). The main benefit of such an approach is the reduction in cycle length and phenotypic expenses. A few validation experiments of marker-assisted recurrent selection with GS have been recently published and a number of others are under way (Asoro et al. 2013; Combs and Bernardo 2013; Massman et al. 2012b). Overall, they confirmed the efficiency and superiority of GS over classical marker-assisted

selection. But they did not consider a case where resources used for GS would be allocated to more phenotyping or increased population sizes.

A number of practical issues remain to be considered for efficient use of marker-assisted recurrent selection. Selfing might be required after a cycle to generate enough seed for crosses in the next cycle, slowing down the scheme. Selection on multiple traits, including high heritability traits such as height and flowering time, should be included to effectively deliver genetic gain. Because prediction accuracy decreases quickly with the cycles of recurrent genomic selection (Long et al. 2011; Muir 2007), the optimal number of cycles is unclear. Further, phenotyping some of candidates, a practice called, updating the training population, will increase the cycle length and costs.

Identifying adequate training populations for those recurrent selection schemes is another area to be investigated. In the classical marker-assisted recurrent selection scheme based on QTL detection, biparental or connected crosses were generated and used as training data (Xu and Crouch 2008). Published marker-assisted recurrent selection schemes with GS have mostly used a similar approach (Combs and Bernardo 2013; Massman et al. 2012b). This approach requires that resources are concentrated on a small fraction of the breeding population. Models using historical data for training models to predict within families would increase the efficiency of GS in recurrent selection. New models need to be developed to address this issue as empirical results (Massman et al. 2012a; Riedelsheimer et al. 2013) indicate that current models have inconsistent prediction accuracies within-families if no member of the family is included in the training population.

Improved use of phenotype with markers

Until recently, it was not realistic to consider genotyping all phenotyped individuals because of the high cost and low throughput of the earlier genotyping technologies. In that context, the use of markers was restricted to small thoroughly phenotyped population subsets.

However, as the cost of whole-genome genotyping has dropped very significantly (Elshire et al. 2011), emphasis should be shifted to maximizing the value of expensive phenotypes (Myles et al. 2009). It is likely that genotyping costs will continue to decrease in the future with the advancement in sequencing technologies. On the contrary, it is unlikely that phenotyping costs will decrease because of increased energy, labor, equipment, and land costs. As a consequence, phenotypes will increasingly be the most valuable asset and molecular markers should be used to extract all possible information useful to selection from these phenotypes. Molecular markers provide a way to estimate the covariance between the performances of individuals. That information can be used to increase accuracy of selection (Endelman and Jannink 2012) to maximize the value of the phenotype data available. This is useful only for traits with low plot mean heritabilities. As replication increases, usefulness of markers decreases, especially if the number of markers is low because the covariance between individuals is poorly estimated. It has been argued that models predicting non-additive effects such as reproducing kernel Hilbert spaces (RKHS) (de Los Campos et al. 2009) are needed, especially as selection candidates approach commercial release. It is the genotypic and not the breeding value that is important to create a successful cultivar. However, as an individual approaches commercial release, the amount of phenotype data available on the individual itself greatly reduces the usefulness of information from relatives, even with non-additive models.

Adequately combining information from markers and phenotype for phenotyped individuals might seem to be a challenge. Early molecular breeding work advocated for an index combining phenotypic performance and marker information (Lande and Thompson 1990). However, because of mixed model optimality properties (Searle et al. 1992), the index weights should be 0 for the phenotype and 1 for the GS prediction. For a detailed demonstration see the supplement in (Endelman et al. 2013). In practice, this means that if a trial or set of trials is analyzed by GBLUP, the genomic estimated breeding values (GEBV) for those individuals contain all the information available in both the markers and the phenotype and should be directly used to make a selection.

New phenotyping strategies are needed

In the context of marker-assisted recurrent selection with GS, (Lorenz et al. 2011) pointed out that with GS, phenotyping is done to train a model, not to directly select. As a consequence, the unit of evaluation is not the individual but the allele. This suggests new phenotyping strategies, maximizing the replication of alleles over the replication of individuals. This would suggest favoring unreplicated experimental designs over more replications of the same genotypes. More generally, when markers are used to help analyze phenotypes, the switch to allele evaluation is not as clear. The unit of selection remains the individual and markers provide a way to use information from relatives to improve accuracy of prediction of individuals' performance. This has a number of consequences for phenotyping strategies. Because trials are analyzed with markers, the assumption used in experimental design that individuals are independent is no longer valid. Early work with simple pedigrees showed that equivalent optimal designs are not equally statistically efficient when individuals are not independent (de S. Bueno Filho and Gilmour 2003). In some cases, an optimal design with related individuals might not be optimal if they are considered to be independent. This

strongly suggests that current experimental designs need to be reassessed for efficiency.

Given simple assumptions on heritabilities, criteria such as the predicted error variance (PEV) can be derived from the mixed model equations, and used for experimental design optimization prior to the experiment (Laloë et al. 1996; Laloë 1993). (Rincent et al. 2012) used those criteria in a maize population to select an optimal subset of individuals to be phenotyped, such that they would best predict those not phenotyped. There are two caveats to this approach. First, it assumes that the covariances among individuals are known when in fact they are estimated. Given this assumption, mixed model properties imply that adding more individuals or observations to the training population is never detrimental to prediction accuracy. The mixed model (e.g., PEV) criteria are then not useful to select an optimal subset of data already available because they will always point to using all data. Only external criteria, such as when there is a budget constraint, might suggest that a subset of data will be better than all data. The last section of this review discusses in more detail the issue with the estimation of the covariance between individual performances. Second, in a breeding context, what matters for breeding gains is the accuracy on both phenotyped and unphenotyped individuals (Endelman et al. 2013) and not only on unphenotyped individuals. Criteria from (Rincent et al. 2012) can be modified for this purpose. (Cullis et al. 2006) also proposed mixed model derived criteria to optimize unreplicated field trials, taking into account the spatial correlation between the residuals. They noted that their method can also be used to take into account relatedness information.

In the private sector, it has been suggested that GS could be used to increase the size of the breeding programs. For example, large doubled-haploid, full-sib families can be developed and a subset phenotyped for predicting the unphenotyped ones. This seems

an appealing strategy on paper but it is not necessarily a good use of resources depending on relative costs of markers, population development, and phenotyping. Current reports in the literature indicate that, depending on relative costs, it is usually more efficient to phenotype all the individuals in a preliminary yield trial but with fewer replications and more environments if all individuals are genotyped (Endelman et al. 2013; Lorenz 2013). (Endelman et al. 2013) also considered a scenario where the genotyping budget is used instead to increase population size or phenotyping, showing that genotyping was not always beneficial depending on the cost of markers and selection intensity.

Efficient selection on multiple traits

The integration of information from different traits for selection purposes requires renewed interest. This is not a new issue and the theory of selection based on indices (Falconer and Mackay 1996) is well studied and developed, but more research is needed to apply them in practice and identify the weights to give to the different traits in the index to simplify selection.

In animal breeding, indices are widely used, probably because most of the information is available at once on many traits with varying degrees of accuracy and has to be used to make a selection decision on very large populations. The use of indices in plants is less common, because selection traditionally occurs on different traits at different times. In wheat, (*Triticum aestivum*, L.) phenotypic selection for example, yield data is usually not available from preliminary trials used to select on plant height, flowering time and agronomic type.

With GS, information on many traits will become available at the same time and selection will be needed on large sets of individuals, necessitating the use of indices to make optimal use of that information.

The use of indices can be more important for GS because GEBVs are not on the same scale as the raw phenotypic data: They include shrinkage accounting for the varying amount of information available for each individual and are centered on zero. Because of this, GEBVs are optimal for truncation selection (Searle et al. 1992 p263-264). However, this is a complication for traits where phenotypic selection is often based on a threshold value determined with a few check individuals or under stabilizing selection. A check individual will have more phenotypic data than most individuals and, as such, its GEBV value will be less shrunken toward the mean of the population than most other individuals. Developing indices might seem a difficult task if economic weights have to be identified for every trait. However, breeders already subjectively select genotypes based on performance for multiple traits, such that the historical data for breeding programs contain this information. (Bernardo 1991) proposed retrospective selection indices describing selection already practiced in a population and quantifying the relative trait weights used intuitively by a breeder. Those indices should be more accessible to plant breeders because they can be calculated with a mixed model analysis of the historical breeding data.

GS enables multi-trait models in practice

Integration of information on multiple traits is needed for selection purposes. But the covariance between traits can also be used to increase prediction accuracy. Figure 1.3 summarizes the different sources of information available for performance prediction.

The theoretical background for multi-trait models has existed for a long time. However, it is seldom deployed in plants because the lack of balance of the data made it difficult, or impossible, to fit when individuals are assumed to be independent. The use of markers to estimate the covariance between individuals greatly simplifies the implementation of multi-trait models. Those multi-trait mixed models should prove

useful to increase the accuracy of selection for traits difficult and expensive to phenotype such as drought tolerance (Calus and Veerkamp 2011; Jia and Jannink 2012). A correlated trait with higher heritability such as yield under non-stress conditions could be used to increase accuracy of selection for yield under drought.

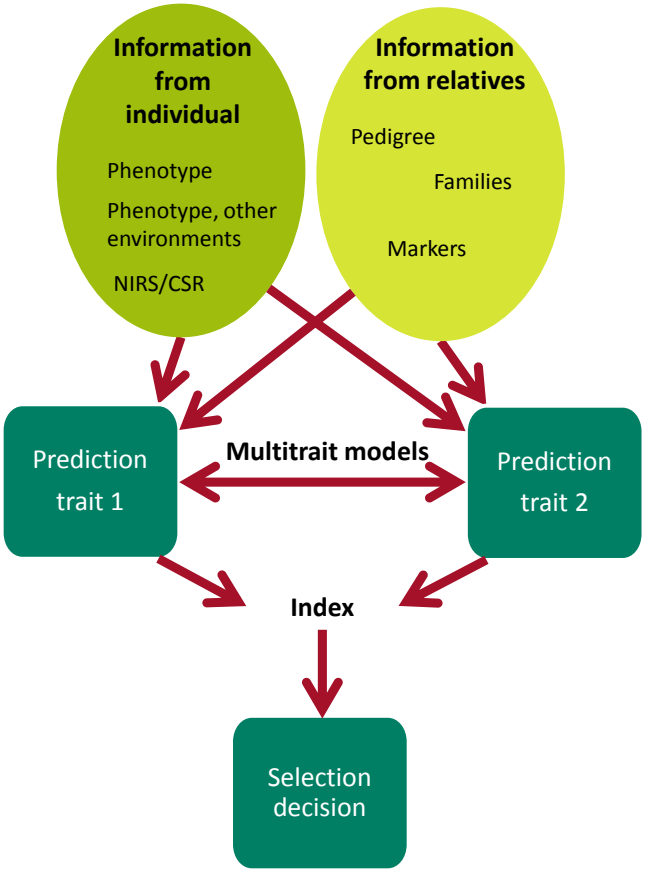


Figure 1.3. Different sources of information for performance prediction and their integration in a breeding program. Arrows indicate the flow of information. NIRS: Near-infrared spectroscopy, CSR: Canopy spectral reflectance.

Another application is where traits of interest are so expensive to phenotype that phenotypic selection uses correlated traits for most of the breeding. For example, malting and baking qualities in cereals, and ethanol yield in maize for biofuels are usually evaluated through near-infrared spectroscopy (NIRS) or through cheap indirect chemical assays. Individuals are tested in industrial conditions, if ever, only in the

very late stages of cultivar development. Multi-trait models could be used to deliver a prediction of end use quality at an early stage by combining marker data and the currently used correlated traits. (Rutkoski et al. 2012) reported higher accuracies for mycotoxin content in wheat, an expensive trait to phenotype with low heritability, by combining markers and simple disease scores.

Similarly, drought tolerance could be predicted using multi-trait GS models combining markers, the available yield under drought data, and inexpensive assays such as canopy spectral reflectance (CSR). Recent empirical results indicate that direct selection for maize yield under drought with GS would be more efficient than indirect phenotypic selection on correlated traits such as anthesis-silking interval (ASI), leaf senescence or leaf chlorophyll content (Ziyomo and Bernardo 2013). However, using markers and correlated traits in a multi-trait model, at the same time should deliver the most gain.

Finally, genotype by environment interactions (G*E) can be analyzed in the multi-trait context by considering performances in different environments as different correlated traits (Falconer 1952). This can be used to increase the accuracy in different target regions (Burgueño et al. 2012; Cullis et al. 2010; Piepho and Möhring 2005).

*Accommodating G*E in GS*

Genotype by environment interactions (G*E) is not a new issue in plant breeding (Cooper and Hammer 1996) but it presents specific opportunities and challenges for GS. G*E is the differential response of individuals to dissimilar environments, potentially leading to change of ranks between individuals (cross-over G*E) which complicate or make impossible the selection of one individual performing best everywhere.

Genomic prediction allows the use of historical phenotypic data to make a selection, thus basing the selection on a broader set of environments than the typical few years of data used by classical phenotypic selection (Heffner et al. 2009). This is beneficial in breeding for stability because even if an individual has not been tested in a specific environment, some of its relatives may have been, allowing estimation of its own performance there.

Underlying the problem of assessing genotype stability, is the issue of correctly sampling and defining the target population of environments (TPE) (Podlich et al. 1999; Tardieu 2012). The TPE is the mixture of environments, defined by both abiotic (e.g., weather and soil) and biotic (e.g., weed and disease) parameters, that are likely to occur in the region where breeding program cultivars will be grown. Genetic gains for performance in the TPE can be impacted, in the presence of G*E, if the data used for selection is not a representative sample of the TPE or if the TPE structure is not accounted for in the analysis. It is likely that not all historical data is relevant for performance in the TPE (Heslot et al. 2013b). Defining the TPE may be more critical with GS than with phenotypic selection: in the latter, data are typically only used to select or discard the specific breeding line on which they were measured. Thus, if a particular year of data is a bad sample of the TPE, it will impact genetic gain for only a short period of time. In the former, in contrast, the unrepresentative data may affect genetic gain over a longer period of time as it will influence marker effect estimates or performance of relatives that, in turn, will affect selection criteria going forward.

GS also opens new ways of analyzing and coping with G*E. As noted previously, analysis of G*E with multi-trait models becomes more tractable with markers by helping to cope with unbalanced data (Burgueño et al. 2012). Considering allele replication rather than individual replication, marker effects in each environment can

be used to cluster environments and identify outliers in highly unbalanced phenotypic data sets (Heslot et al. 2013b).

GS provides an opportunity to integrate environmental covariates (e.g., climate data) to predict G*E deviations for unobserved environments (Heslot et al. 2013a). Genome-wide marker effects can be considered as a function of environmental covariates that are estimated using GS methods. This approach can in turn allow prediction of individual stability, identification of important stresses and investigation of the TPE structure that is critical for breeding strategies (Podlich et al. 1999).

Need for improved prediction models

A lot of attention in GS was initially devoted to statistical models (Gianola et al. 2009) but current GS models behave similarly on empirical data (Heslot et al. 2012). The GBLUP model seems efficient in most situations but additional research would be beneficial as described below.

First, major QTLs are known for a number of traits in plants. Applying GS in that context might seem problematic. However, (Bernardo 2013) showed in simulations that it is beneficial to fit the known QTLs as fixed effects only when they each explain more than 10% of the genetic variance. Because their simulations assume that major QTLs are known and in complete linkage disequilibrium with a marker, the practical threshold should probably be higher. Overall, this indicates that GS should be effective for most traits, even when large QTLs are present and without the need for identification or special treatment of the large QTLs.

A more subtle shortcoming of GBLUP was recently identified. Optimality of GBLUP is based on knowledge of the true covariance between individuals. The true covariance between individuals for a given trait depends on the relationship at causal loci and not on the whole genome relationship (Endelman and Jannink 2012). Hill and

Weir (2011) and de los Campos et al. (2013) showed, that even for a complex trait at very high marker density, the whole genome relationship does not approximate well the relationship at causal loci for distantly related individuals. This is a likely explanation of why prediction across breeds in dairy cattle does not seem to work (Erbe et al. 2012) with GBLUP and that sometimes adding more individuals to the training set actually decreases prediction accuracy (Habier et al. 2013; Riedelsheimer et al. 2013).

If the true covariance was used for the analysis, accuracy should not decrease with increasing training population size. The apparent difficulty to adequately predict within families when no individuals of the family are phenotyped (Massman et al. 2012a) is also likely linked to the issue of poor relationship approximation at causal loci for distantly related individuals with GBLUP.

If the covariance is not well estimated, adding more individuals to the training population can be detrimental, but training population mixed model optimization criteria assumes that covariance is known. As a consequence, adding more individuals is always beneficial for those optimization criteria. For example in dairy cattle, predicted accuracies based on mixed model criteria were good predictors of observed accuracies for within breeds models but not for between breeds models (Hayes et al. 2009a). This observation reveals an interesting connection between training population design and covariance estimation. If the covariance is better estimated, more individuals can be useful in the training population.

There are two potential strategies to overcome this issue. One would be to use complete sequencing (Meuwissen 2010) so that the causal loci would be in the data and variable selection methods could be used to identify them. However, empirical, simulation, and theoretical studies put into doubt the value of this approach (Gianola 2013; Wimmer et al. 2013) because of the large excess of marker effects to be

estimated compared to the number of observations. Nevertheless, it cannot be ruled out that variable selection methods could be useful.

To help identify the causal loci for prediction purposes, other sources of information can be used such as p-values from GWAS on different datasets (de los Campos et al. 2013) or from other genomic sources. Prior information about the potential effect of a polymorphism in coding sequence in humans can be obtained with the use of software such as Polyphen (Adzhubei et al. 2010). Similarly, polymorphisms identified to be in regulatory regions can be given a higher prior probability of contributing to the trait than polymorphisms in non-coding non regulatory regions using results from the ENCODE project in humans (Maurano et al. 2012). Similar approaches could be developed in plants to derive informative priors for marker effects for genomic prediction purposes.

Another avenue of research would be to develop new covariance estimators. If two individuals are distantly related based on pedigree, the observed kinship coefficient is a poor estimate of the covariance at causal loci. Whereas the same observed kinship coefficients but between two highly related individuals is a good estimate of the covariance at causal loci (Hill and Weir 2011; de los Campos et al. 2013). This could be taken into account by shrinking certain coefficients of the kinship matrix.

Conclusion

GS provides tremendous opportunities to increase genetic gain in plant breeding. Early empirical and simulation results are promising but for GS to work, a systems perspective that considers the problem of resource allocation is needed. It is also important to understand that markers can be used to improve accuracy of selection even for phenotyped individuals. Use of markers for that purpose suggests a number of new ways to improve phenotyping strategies in terms of experimental design and multi-trait models. GS also provides new ways to analyze and deal with G*E. Finally, more work is needed to develop better prediction models for distantly related individuals. These are my take home messages and directions I believe will be fruitful for future research.

REFERENCES

Adzhubei IA, Schmidt S, Peshkin L, et al. (2010) A method and server for predicting damaging missense mutations. Nat Methods doi: 10.1038/nmeth0410-248

Asoro FG, Newell MA, Beavis WD, et al. (2013) Comparison of genomic, marker-assisted, and pedigree-BLUP selection methods to increase β-glucan concentration in elite oat germplasm. Crop Sci. doi: 10.2135/cropsci2012.09.0526

Bernardo R (2008) Molecular markers and selection for complex traits in plants: Learning from the last 20 years. Crop Sci doi: 10.2135/cropsci2008.03.0131

Bernardo R (2009) Genomewide selection for rapid introgression of exotic germplasm in maize. Crop Sci doi: 10.2135/cropsci2008.08.0452

Bernardo R (1991) Retrospective index weights used in multiple trait selection in a maize breeding program. Crop Sci doi: 10.2135/cropsci1991.0011183X003100050020x

Bernardo R (2013) Genomewide selection when major genes are present. doi: 10.2135/cropsci2013.05.0315

Bernardo R, Yu J (2007) Prospects for genomewide selection for quantitative traits in maize. Crop Sci doi: 10.2135/cropsci2006.11.0690

Burgueño J, de los Campos G, Weigel K, Crossa J (2012) Genomic prediction of breeding values when modeling genotype × environment interaction using pedigree and dense molecular markers. Crop Sci doi: 10.2135/cropsci2011.06.0299

Calus MPL, Veerkamp RF (2011) Accuracy of multi-trait genomic selection using different methods. Genet Sel Evol doi: 10.1186/1297-9686-43-26

Combs E, Bernardo R (2013) Genomewide selection to introgresssemidwarf maize germplasm into U.S. corn belt inbreds. Crop Sci doi: 10.2135/cropsci2012.11.0666

Cooper M, Hammer GL editors (1996) Plant adaptation and crop improvement. CAB International, Wallingford, UK

Crossa J, de Los Campos G, Pérez P, et al. (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics doi: 10.1534/genetics.110.118521

Cullis BR, Smith AB, Beeck CP, Cowling WA (2010) Analysis of yield and oil from a series of canola breeding trials. Part II. Exploring variety by environment interaction using factor analysis. Genome doi: 10.1139/G10-080

Cullis BR, Smith AB, Coombes N (2006) On the design of early generation variety trials with correlated data. J Agric Biol Environ Stat doi: 10.1198/108571106X154443

Elshire RJ, Glaubitz JC, Sun Q, et al. (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One doi: 10.1371/journal.pone.0019379

Endelman JB, Atlin GN, Beyene Y, et al. (2013) Optimal design of preliminary yield trials with genome-wide markers. Crop Sci. doi: 10.2135/cropsci2013.03.0154

Endelman JB, Jannink J-L (2012) Shrinkage estimation of the realized relationship matrix. Genes, Genomes, and Genomics doi: 10.1534/g3.112.004259

Erbe M, Hayes BJ, Matukumalli LK, et al. (2012) Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J Dairy Sci doi: 10.3168/jds.2011-5019

Falconer DS (1952) The Problem of Environment and Selection. Am Nat 86:293–298.

Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics, 4th ed. Pearson, Prentice Hall, Harlow, UK

Gianola D (2013) Priors in whole-genome regression: The bayesian alphabet returns. Genetics. doi: 10.1534/genetics.113.151753

Gianola D, de Los Campos G, Hill WG, et al. (2009) Additive genetic variability and the Bayesian alphabet. Genetics doi: 10.1534/genetics.109.103952

Goddard ME (2009) Genomic selection: prediction of accuracy and maximisation of long term response. Genetica doi: 10.1007/s10709-008-9308-0

Habier D, Fernando RL, Dekkers JCM (2007) The impact of genetic relationship information on genome-assisted breeding values. Genetics doi: 10.1534/genetics.107.081190

Habier D, Fernando RL, Garrick DJ (2013) Genomic-BLUP decoded: A look into the black box of genomic prediction. Genetics doi: 10.1534/genetics.113.152207

Hayes BJ, Bowman PJ, Chamberlain AJ, et al. (2009a) Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genet Sel Evol doi: 10.1186/1297-9686-41-51

Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009b) Invited review: Genomic selection in dairy cattle: progress and challenges. J Dairy Sci doi: 10.3168/jds.2008-1646

Heffner EL, Jannink J-L, Iwata H, et al. (2011) Genomic selection accuracy for grain quality traits in biparental wheat populations. Crop Sci. doi: 10.2135/cropsci2011.05.0253

Heffner EL, Lorenz AJ, Jannink J-L, Sorrells ME (2010) Plant breeding with genomic selection: gain per unit time and cost. Crop Sci doi: 10.2135/cropsci2009.11.0662

Heffner EL, Sorrells ME, Jannink J-L (2009) Genomic selection for crop improvement. Crop Sci doi: 10.2135/cropsci2008.08.0512

Henderson CR (1984) Applications of linear models in animal breeding, University of Guelph, Guelph, Ontario

Heslot N, Akdemir D, Sorrells ME, Jannink J-L (2013a) Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. Theor Appl Genet. doi: 10.1007/s00122-013-2231-5

Heslot N, Jannink J-L, Sorrells ME (2013b) Using genomic prediction to characterize environments and optimize prediction accuracy in applied breeding data. Crop Sci doi: 10.2135/cropsci2012.07.0420

Heslot N, Rutkoski JE, Poland J, et al. (2013c) Impact of marker ascertainment bias on genomic selection accuracy and estimates of genetic diversity. PLoS One doi: 10.1371/journal.pone.0074612

Heslot N, Yang H-P, Sorrells ME, Jannink J-L (2012) Genomic selection in plant breeding: A comparison of models. Crop Sci doi: 10.2135/cropsci2011.06.0297

Hill WG, Weir BS (2011) Variation in actual relationship as a consequence of Mendelian sampling and linkage. Genet Res (Camb) doi: 10.1017/S0016672310000480

Jannink J-L (2010) Dynamics of long-term genomic selection. Genet Sel Evol doi: 10.1186/1297-9686-42-35

Jia Y, Jannink J-L (2012) Multiple trait genomic selection methods increase genetic value prediction accuracy. Genetics doi: 10.1534/genetics.112.144246

König S, Simianer H, Willam A (2009) Economic evaluation of genomic breeding programs. J Dairy Sci doi: 10.3168/jds.2008-1310

Laloë D (1993) Precision and information in linear models of genetic evaluation. Genet Sel Evol 25:556–576.

Laloë D, Phocas F, Ménissier F (1996) Considerations on measures of precision and connectedness in mixed linear models of genetic evaluation. Genet Sel Evol. doi: 10.1051/gse:19960404

Lande R, Thompson R (1990) Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124:743–756.

Lillehammer M, Meuwissen THE, Sonesson AK (2013) A low-marker density implementation of genomic selection in aquaculture using within-family genomic breeding values. Genet Sel Evol doi: 10.1186/1297-9686-45-39

Long N, Gianola D, Rosa GJM, Weigel K (2011) Long-term impacts of genome-enabled selection. J Appl Genet. doi: 10.1007/s13353-011-0053-1

Lorenz AJ (2013) Resource allocation for maximizing prediction accuracy and genetic gain of genomic selection in plant breeding: a simulation experiment. G3 doi: 10.1534/g3.112.004911

Lorenz AJ, Chao S, Asoro FG, et al. (2011) Genomic selection in plant breeding : knowledge and prospects. Adv Agron doi: 10.1016/B978-0-12-385531-2.00002-5

Lorenz AJ, Smith KP, Jannink J-L (2012) Potential and optimization of genomic selection for fusarium head blight resistance in six-row barley. Crop Sci doi: 10.2135/cropsci2011.09.0503

Lorenzana RE, Bernardo R (2009) Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theor Appl Genet doi: 10.1007/s00122-009-1166-3

De Los Campos G, Gianola D, Rosa GJM (2009) Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. J Anim doi: 10.2527/jas.2008-1259

De los Campos G, Vazquez AI, Fernando RL, et al. (2013) Prediction of complex human traits using the genomic best linear unbiased predictor. PLoS Genet doi: 10.1371/journal.pgen.1003608

Massman JM, Gordillo A, Lorenzana RE, Bernardo R (2012a) Genomewide predictions from maize single-cross data. Theor Appl Genet. doi: 10.1007/s00122-012-1955-y

Massman JM, Jung H-JG, Bernardo R (2012b) Genomewide selection versus marker-assisted recurrent selection to improve grain yield and stover-quality traits for cellulosic ethanol in maize. Crop Sci. doi: 10.2135/cropsci2012.02.0112

Maurano MT, Humbert R, Rynes E, et al. (2012) Systematic localization of common disease-associated variation in regulatory DNA. Science doi: 10.1126/science.1222794

Meuwissen THE (2010) Use of whole genome sequence data for QTL mapping and genomic selection. Proc. 9th World Congr. Genet. Appl. to Livest. Prod.

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–29.

Moreau L, Charcosset A, Gallais A (2004) Experimental evaluation of several cycles of marker-assisted selection in maize. Euphytica doi: 10.1023/B:EUPH.0000040508.01402.21

Muir WM (2007) Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. J Anim Breed Genet doi: 10.1111/j.1439-0388.2007.00700.x

Myles S, Peiffer JA, Brown PJ, et al. (2009) Association mapping: critical considerations shift from genotyping to experimental design. Plant Cell doi: 10.1105/tpc.109.068437

Piepho HP, Möhring J (2005) Best linear unbiased prediction of cultivar effects for subdivided target regions. Crop Sci 45:1151. doi: 10.2135/cropsci2004.0398

Piepho HP, Möhring J, Melchinger AE, Büchse A (2007) BLUP for phenotypic selection in plant breeding and variety testing. Euphytica doi: 10.1007/s10681-007-9449-8

Podlich DW, Cooper M, Basford KE (1999) Computer simulation of a selection strategy to accommodate genotype-environment interactions in a wheat recurrent selection programme. Plant Breed doi: 10.1046/j.1439-0523.1999.118001017.x

Riedelsheimer C, Endelman JB, Stange M, et al. (2013) Genomic predictability of interconnected bi-parental maize populations. Genetics. doi: 10.1534/genetics.113.150227

Rincent R, Laloë D, Nicolas S, et al. (2012) Maximizing the reliability of genomic delection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (Zea mays L.). Genetics doi: 10.1534/genetics.112.141473

Rutkoski JE, Benson J, Jia Y, et al. (2012) Evaluation of genomic prediction methods for fusarium head blight resistance in wheat. Plant Genome J 5doi: 10.3835/plantgenome2012.02.0001

De S. Bueno Filho JS, Gilmour SG (2003) Planning incomplete block experiments when treatments are genetically related. Biometrics doi: 10.1111/1541-0420.00044

Schaeffer LR (2006) Strategy for applying genome-wide selection in dairy cattle. J Anim Breed Genet doi: 10.1111/j.1439-0388.2006.00595.x

Searle SR, Casella G, McCulloch. CE (1992) Variance components. John Wiley, Hoboken

Servin B, Martin OC, Mézard M, Hospital F (2004) Toward a theory of marker-assisted gene pyramiding. Genetics doi: 10.1534/genetics.103.023358

Smith AB, Cullis BR, Thompson R (2005) The analysis of crop cultivar breeding and evaluation trials: an overview of current mixed model approaches. J Agric Sci doi: 10.1017/S0021859605005587

Stuber CW, Goodman MM, Moll RH (1982) Improvement of yield and ear number resulting from selection at allozyme Loci in a maize population. Crop Sci doi: 10.2135/cropsci1982.0011183X002200040010x

Tanksley SD, Young ND, Paterson AH, Bonierbale MW (1989) RFLP mapping in plant breeding: new tools for an old science. Nat Biotechnol doi: 10.1038/nbt0389-257

Tardieu F (2012) Any trait or trait-related allele can confer drought tolerance: just design the right drought scenario. J Exp Bot 63:25–31. doi: 10.1093/jxb/err269

Tribout T, Larzul C, Phocas F (2013) Economic aspects of implementing genomic evaluations in a pig sire line breeding scheme. Genet Sel Evol doi: 10.1186/1297-9686-45-40

Wimmer V, Lehermeier C, Albrecht T, et al. (2013) Genome-wide prediction of traits with different genetic architecture through efficient variable selection. Genetics doi: 10.1534/genetics.113.150078

Xu Y, Crouch JH (2008) Marker-assisted selection in plant breeding: From publications to practice. Crop Sci doi: 10.2135/cropsci2007.04.0191

Zhong S, Jannink J-L (2007) Using quantitative trait loci results to discriminate among crosses on the basis of their progeny mean and variance. Genetics doi: 10.1534/genetics.107.075358

Ziyomo C, Bernardo R (2013) Drought tolerance in maize: indirect selection through secondary traits versus genomewide selection. Crop Sci doi: 10.2135/cropsci2012.11.0651

CHAPTER 2

GENOMIC SELECTION IN PLANT BREEDING: A COMPARISON OF MODELS[2]

Abstract

Simulation and empirical studies of genomic selection (GS) show accuracies sufficient to generate rapid genetic gains. However, with the increased popularity of GS approaches, numerous models have been proposed and no comparative analysis is available to identify the most promising ones. Using eight wheat (*Triticum aestivum* L.), barley (*Hordeum vulgare* L.), *Arabidopsis thaliana* (L.) Heynh., and maize (*Zea mays* L.) datasets, the predictive ability of currently available GS models along with several machine learning methods was evaluated by comparing accuracies, the genomic estimated breeding values (GEBVs), and the marker effects for each model. While a similar level of accuracy was observed for many models, the level of overfitting varied widely as did the computation time and the distribution of marker effect estimates. My comparisons suggested that GS in plant breeding programs could be based on a reduced set of models such as the Bayesian Lasso, weighted Bayesian shrinkage regression (wBSR, a fast version of BayesB), and random forest (RF) (a machine learning method that could capture nonadditive effects). Linear combinations of different models were tested as well as bagging and boosting methods, but they did not improve accuracy. This study also showed large differences in accuracy between subpopulations within a dataset that could not always be explained by differences in phenotypic variance and size. The broad diversity of empirical datasets tested here adds evidence that GS could increase genetic gain per unit of time and cost.

[2] Heslot N, Yang H-P, Sorrells ME, Jannink J-L (2012) Genomic selection in plant breeding: A comparison of models. Crop Sci doi: 10.2135/cropsci2011.06.0297

Abbreviations

BL, Bayesian Lasso; BRR, Bayesian ridge regression; CAP, Coordinated Agricultural Project; E-Bayes, empirical Bayes; EM, expectation maximization; *Fst*, the *F* statistics quantifying the differences in allele frequency among subpopulations; GEBV, genomic estimated breeding value; GS, genomic selection; LD, linkage disequilibrium; MCMC, Markov chain Monte Carlo; NNET, neural network; PCA, principal component analysis; QTL, quantitative trait loci/locus; RF, random forest; RKHS, reproducing kernel Hilbert space; RR-BLUP, random regression best linear unbiased predictor; SVM, support vector machine; SVR, support vector regression; wBSR, weighted Bayesian shrinkage regression

Introduction

The genomic selection (GS) concept encompasses a broad range of methods. Their common feature is the ability to estimate breeding values for quantitative traits based on whole genome genotypes through the simultaneous estimation of marker effects in a single step. This concept was first proposed by (Meuwissen et al. 2001) with several new statistical models. Since then, further models have been proposed. Simulations and empirical studies have demonstrated that GS can greatly accelerate the breeding cycle, maintain genetic diversity within the breeding programs, and increase genetic gain beyond what is possible with phenotypic selection or quantitative trait loci (QTL) approaches. Nevertheless, it is important to identify the best methods and statistical procedures for using high-throughput molecular marker technologies and previously available phenotypic records to accelerate genetic gains per unit of time and cost. Several recent reviews are available on GS in plant breeding, in particular (Heffner et al. 2009; Jannink et al. 2010; Lorenz et al. 2011; Xu and Hu 2010) .

There are few extensive studies of the comparative predictive ability of the proposed models in plants or in animals. (Lorenzana and Bernardo 2009) showed that, in the case of biparental populations, the predictive ability of the models they tested (ridge regression and empirical Bayes [E-Bayes] (Xu 2007)) was fairly similar. (Heffner et al. 2011) compared several models for predictive ability in a multiparental wheat (*Triticum aestivum* L.) population. (Crossa et al. 2010) focused on Bayesian Lasso (BL) and reproducing kernel Hilbert space (RKHS) models to evaluate GS for wheat and maize (*Zea mays* L.) improvement. My objective in this study was to thoroughly compare all the models published to date, along with several machine learning procedures not previously evaluated for GS, using the same evaluation methods on several species, traits, and datasets. In addition, none of the above cited model comparison studies measured the level of overfitting in each model, which is also an important factor to quantify. It should also be emphasized that for a given level of accuracy, models use different assumptions on QTL effect distributions resulting in different marker effect distributions. Therefore, even if they have the same predictive ability, two models will likely give different genomic estimated breeding values (GEBVs) and exert different selection pressures along the genome. My goal was to identify the most promising models, provide some recommendations for the implementation of GS approaches in breeding programs, and obtain empirical evidence of model similarities and dissimilarities.

Materials and methods

Phenotypic and Genotypic Data

Eight datasets of different origins were used (Table 2.1) including two published datasets previously used to test GS models for *Arabidopsis thaliana* (L.) Heynh. (Bay × Sha [Bay-0 × Shahdara]) (Lorenzana and Bernardo, 2009) and wheat (Wheat CIMMYT) (Crossa et al. 2010). The Wheat Cornell dataset used is a subset of the dataset used in (Heffner et al. 2011). The Barley Coordinated Agricultural Project (CAP) dataset was from the Barley Coordinated Agricultural Project (2011). All other datasets were provided by Limagrain Europe (Chappes, France). For both maize datasets, phenotype data were obtained from a testcross to a Limagrain Europe inbred. Several types of markers were used. Diversity array technology markers (DArT) markers are dominant and single nucleotide polymorphism (SNP) and simple sequence repeat (SSR) are codominant. Missing marker data were imputed as the mean of the nonmissing data at the level of each marker.

Table 2.1. Dataset origins and details.

†By "panel" I mean a group of mostly unrelated lines.

‡SNP, single nucleotide polymorphism.

§CAP, Coordinated Agricultural Project.

¶SSR, simple sequence repeat.

#INRA, Institut National de la Recherche Agronomique (Paris, France).

††BLUE, best linear unbiased estimator.

‡‡GCA, general combining ability.

§§DArT, diversity array technology markers (Triticarte Pty. Ltd. Canberra, Australia).

Name	Crop	Type	Traits	Number of genotypes	Phenotypic data	Number and type of marker	Origin
Barley 1	Spring barley	Panel†, elite breeding lines	Yield	761	Three years, eight trial locations per year on average, across mainland Europe, nine replicates per genotype on average	338 SNP‡	Limagrain Europe (Chappes, France)
Barley CAP§	Spring barley	Panel with structure	Betaglucan content	911	Three years, unbalanced data from five locations per year. Lines per trial ranged from 22 to 96.	2146 SNP	Barley CAP project
Bay × Sha (Bay-0 × Shahdara)	Arabidopsis thaliana (L.) Heynh.	Biparental population from two ecotypes, Bay-0 and Shadara	Flowering time under short day length, dry matter under nonlimiting or limiting conditions	422	Data available from the Study of the Natural Variation of *Arabidopsis thaliana* website (INRA, 2007)	69 SSR¶	INRA# (France) (Lorenzana and Bernardo 2009; Loudet et al. 2002)
Panel maize	Elite maize	Panel, elite lines (one heterotic group)	Yield and moisture content	332	BLUE†† of GCA‡‡ for both traits, tested in 2009 in northern Italy.	355 SNP	Limagrain Europe
Diallel maize	Elite maize	Partial diallel, elite lines (one heterotic group) (six parents and five crosses)	Yield and moisture content	370	BLUE of GCA for both traits, tested in 2009 in northern Italy.	319 SNP	Limagrain Europe
Wheat CIMMYT	Spring wheat	Panel	Yield measured in four different target environments	599	Environments were grouped into four target sets (E1–E4)	1279 DArT§§	CIMMYT (Crossa et al., 2010)
Wheat Cornell	Winter wheat	Panel with family structure	Yield and heading date	374	One year (2009), two locations in New York state and two replicates per genotypes	1158 DArT	Cornell University (Heffner et al. 2011)
Wheat diallel	Winter wheat	Partial diallel, elite lines (eight crosses and five parents)	Yield, plant height, and thousand kernel weight	551	Three years, two to five trial locations per year in France, 18 replicates per genotype on average	319 SNP	Limagrain Europe

Models tested

Eleven GS models were used to estimate the genetic value of individuals. Random regression best linear unbiased predictor (RR-BLUP), also named ridge regression, was used with either a grid search over the shrinkage parameter λ or an estimation of the level of shrinkage using a mixed model approach with the "emma" R package (Kang et al. 2008). I also used the Bayesian ridge regression (BRR) as implemented in the R package "BLR" (Pérez et al. 2010). The model is of the form $y = \mu + X\beta + \varepsilon$ where y is the trait value, μ is the population mean, X is the marker design matrix, β is the vector of marker effects, and the error term, ε, is assumed to be normally distributed with mean and variance equal to 0 and σ^2. The estimator of β is $\left(X'X + \lambda I\right)^{-1} X'y$. This estimator can be expressed as: $\arg\min_{\beta}\left(\|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2\right)$ with the notation $\|\beta\|_2 = \left(\sum_i \beta_i^2\right)^{1/2}$ used for the L_2 or Euclidean norm. The notation $\arg\min_{\beta}$ refers to the determination of coefficients β minimizing the expression inside the brackets. Random regression best linear unbiased predictor assumes all markers have a common variance (Meuwissen et al. 2001) and therefore shrinks equally for each marker effect. Bayesian ridge regression makes the same assumptions as RR-BLUP but the level of shrinkage is estimated with a Bayesian hierarchical model.

In the case of the BL (de los Campos et al. 2009; Park and Casella 2008; Yi and Xu 2008), the shrinkage is marker specific and dependent on a regularization parameter λ. The estimator of β is $\arg\min_{\beta}\left(\|y - X\beta\|_2^2 + \lambda\|\beta\|_1\right)$ thus illustrating the similarities between the BL and RR-BLUP. In both cases, $\|y - X\beta\|_2^2$ is a sum of squares penalty while the remaining term is a penalty function promoting sparseness. In the BL, this function is based on the L_1 norm (also named Taxicab or Manhattan norm)

$\|\beta\| = \sum_i |\beta_i|$ while in RR-BLUP it is based on the L_2 norm as described above. The BL produces stronger shrinkage of regression coefficients that are close to zero and less shrinkage of those with large absolute values, leading to a sparse model, whereas RR-BLUP shrinks more strongly the regression coefficients with a large value. For the ridge regression, there are several possibilities to determine λ as described above. In the Bayesian version of the Lasso, each marker effect β_j is assigned a normal prior of mean 0 and variance σ_j^2 . Each follows an independent exponential prior with parameter $\lambda^2/2$. A Gamma prior is further assigned to λ. It is important to note a fundamental difference between the BL and the lasso: The Bayesian version does not select variables by assigning coefficients to 0 as does the non-Bayesian version. For both BRR and BL I used the default prior parameters provided in Pérez et al. (2010) with 60,000 iterations and the first 10,000 iterations were discarded as burn-in.

The elastic net (Zou and Hastie 2005) relies on a combination of both the L_1 norm (lasso) and L_2 norm penalties (ridge regression). The estimator of β is $(1+\lambda_2)\arg\min_\beta\left(\|y - X\beta\|_2^2 + \lambda_2\|\beta\|_2^2 + \lambda_1\|\beta\|_1\right)$ with λ_1 and λ_2 shrinkage parameters. This double regularization generates a sparse model through the L_1 norm penalty and the L_2 part removes the limitation on the number of selected variables, encourages grouping effects, and stabilizes the L_1 path. This model was implemented using the R package glmnet (Friedman et al. 2010). This implementation performs a coordinate descent search for the lasso parameter. In addition I performed a grid search via cross-validation for the other shrinkage parameter controlling the relative amount of L_1 and L_2 penalties. The model with the minimum MSE was selected.

BayesB and its relative, BayesA, (Meuwissen et al., 2001) relax the assumption of common variance across marker effects made by RR-BLUP. The prior for marker effect j is a mixture distribution with β_j equal to 0 with probability π and, with

probability $1-\pi$, β_j is sampled from a normal distribution with mean 0 and variance σ_j^2. Finally, σ_j^2 is sampled from a scaled inverse χ^2 with degrees of freedom ν and scale S^2. In the case of the original BayesB publication, π was set to 0.95. The BayesB model reduces to BayesA for $\pi = 0$. However, the computational demand of those original models limits their implementation even though simulation studies (Habier et al. 2007) stress their advantages over RR-BLUP. For this reason, BayesB was tested only on the smaller datasets (Barley 1 [Limagrain Europe, Chappes, France], Bay × Sha, Diallel maize [Limagrain Europe], and Panel maize [Limagrain Europe]) with two chains of 10,000 iterations and 1000 for burn-in.

The weighted Bayesian shrinkage regression (wBSR) method (Hayashi and Iwata 2010) is an expectation maximization (EM) algorithm for the BayesB model (Meuwissen et al., 2001). Preliminary testing of this model revealed that the initial convergence parameter set up for the algorithm was not adequate for some datasets and would generate unstable results. The authors of this model provided an updated version that allows the user to set the convergence parameter. The high computational efficiency of this algorithm allows a complete grid search to be performed on the prior parameters. The prior parameters searched were ν, the degree of freedom, and S^2, the scale parameter of the scaled inverse χ^2 distribution of the marker effect variance prior and π. Six hundred triplets of prior parameters were tested for each dataset using a 10-fold cross-validation. The range of the grid search for the scale prior parameter was chosen following (Gianola et al. 2009).

BayesCπ (Lorenz et al. 2010) assumes a common marker effect variance for all markers with nonzero effects, but rather than using a fixed π , it estimates π. The model was fitted with a single chain of 10,000 iterations, the first 1000 being

discarded as burn-in. For Bayesian models the Markov chain Monte Carlo (MCMC) algorithm was used to obtain the posterior parameters and visually checked for convergence using the R package "coda" (Plummer et al. 2006).

Empirical Bayes (E-Bayes) (Xu, 2007) is a differential parameter shrinkage method for an oversaturated regression model. The original model was intended to incorporate linear combinations of all additive and pairwise epistatic effects among markers. As in BayesA, the prior for each β_j is assumed to follow a normal distribution with mean 0 and variance σ_j^2. The marker variance parameter is further assumed to be inverse χ^2 distributed with degree of freedom τ and scale parameter ω. However, the E-Bayes algorithm does not require MCMC samplings to estimate the variance parameters. Instead a maximization algorithm is used to reduce computation time. The full model including additive and all pairwise epistatic effects contained too many effects. I therefore tested this model only with additive effects for all datasets. I optimized the model prediction by grid-searching multiple combinations of parameters (τ and ω) for each dataset. The parameter space tested ranged from –2 to –0.5 for τ and from 0 to 0.1 for ω.

The RKHS approach first uses a kernel function to convert the marker dataset into a set of distances between pairs of observations that results in a square matrix to be used in a linear model. Because RKHS regression does not assume linearity it might better capture nonadditive effects. The model can be formulated as $y = W\mu + K_h\alpha + \varepsilon$ where μ is a vector of fixed effects and ε is a vector of random residuals.The parameters α and ε are assumed to have independent prior distributions $\alpha \sim N\left(0, K_h\sigma_\alpha^2\right)$ and $\varepsilon \sim N\left(0, I\sigma_e^2\right)$, respectively. Matrix K_h depends on a reproducing kernel function with a smoothing parameter h, which measures the "genomic distance" between genotypes and can be interpreted as a correlation matrix. Parameter h controls the rate of decay of the correlation between genotypes.

Given h the RKHS regression is the same as a standard mixed-effects linear model. The mutual exchange of information between α-coefficients due to the nontrivial correlation structure induced by K_h is similar to the exchange of information between relatives induced by the genetic additive relationship matrix in the classical additive genetic model.

The kernel function I tested is a Gaussian kernel, $K_h\left(x_i, x_j\right) = \exp\left(-hd_{ij}\right) = \exp\left(-\theta d_{ij} / k\right)$, where d_{ij} is a marker-based distance between two individuals i and j. To decide which parameter combinations to use for optimal predictions for RKHS regression, I tested two distance methods built in R (R Development Core Team 2010) to compute genetic distance: the squared Euclidean and the squared Manhattan distances. I also tested different values of θ and k. The θ values tested ranged from 0.1 to 10. The k value tested include (i) d_{median}, the sample median of d_{ij}, (ii) $d^2_{\max}$, the maximum value of squared distance of d_{ij}, and (iii) m, the number of markers genotyped. Only the Barley CAP dataset was used to optimize parameters for RKHS regression. For all other datasets, I used the same parameter combination that was optimized for the Barley CAP dataset $\theta / k = 2 / d_{median}$ with d_{ij} based on the Manhattan distance to predict trait values.

Machine-learning methods such as random forest (RF) regression (Breiman 2001), support vector regression (SVR) (Drucker et al. 1997), and artificial neural networks (Gardner and Dorling 1998) have been widely used in research and industrial settings. They could also be useful in the prediction of breeding values (González-Recio and Forni 2011; Moser et al. 2009) and identification of causal polymorphisms (Bureau et al. 2005). Since those methods are nonparametric and the underlying theory behind

them is quite different from the linear model for GS approaches described above, they may be able to capture different relationships between markers and phenotypes.

A RF is a collection of classification or regression trees grown on bootstrap samples of observations using a random subset of predictors to define the best split at each node. Different variables are used at each split in different trees. The RF prediction for an observation is computed by averaging the predictions over trees for which the given observation was not used to build the tree. This model was implemented using the R package "RandomForest" (Liaw and Wiener 2002). I used the default setting of the function except for the number of trees, which was set to 1000, and the minimum size of terminal node as 50, as suggested by preliminary testing. I used the tuning function provided to optimize the number of variables randomly sampled at each split for each trait. This model will be referred to as support vector machine (SVM).

Support vector regression (Smola and Schölkopf 2004) uses linear models to implement nonlinear regression by mapping the input space (the marker dataset) to a feature space of a different dimension (lower in the case of GS) using a nonlinear kernel function followed by linear regression in this feature space. The SVR simultaneously minimizes an objective function that accounts for both model complexity and the error in the training data. This model was implemented using the R package "e1071" (Dimitriadou et al. 2011). A linear kernel was used along with an epsilon-insensitive loss function. This means that during the model fitting, all the error up to the epsilon level is simply discarded from the model. A tuning function was used to optimize the level of epsilon and the cost parameter that weights the relative contribution of error and model complexity to the objective function.

The artificial neural network is a very broad class of models inspired by the structure and functions of biological neural networks. It has been demonstrated that the multilayer perceptron, a particular case of neural network, can be trained to approximate virtually any smooth, measurable function (Hornik 1989).The multilayer perceptron is a system of simple interconnected neurons or nodes.

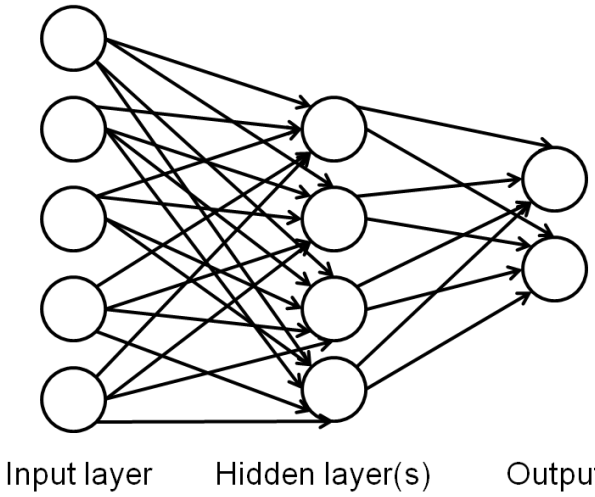


Figure 2.1. A generic feed-forward neural network with a single hidden layer.

Each node sums its inputs multiplied by weights w_{ij}, linking nodes i and j and adding a node specific constant, the bias α_j. The output is then produced by applying an activation function f_i. This activation function can be linear or nonlinear. The system is defined by an input layer with one neuron per input variable x_i (here for each of the N markers), which sends the data to intermediate hidden layers, and by an output layer made of one neuron per output variable y_k that receives input from the last hidden layer. Here, the output layer is made of a single neuron to output the GEBVs. Figure 2.1 gives a general example of a neural network. This graphical example of a network is equivalent to the following function from input to output, with the subscript k indexing the output variables and U the number of nodes in the hidden layer.

$$y_k = f_k\left[\alpha_k + \sum_j^U w_{jk} f_j\left(\alpha_j + \sum_i^N w_{ij} x_i\right)\right]$$

The fitting of a neural network is thus controlled by the number of hidden layers, the number of neurons per hidden layer, the activation function, and the weights of each connection. The training procedure of such a neural network implies the determination of the individual weights and several techniques are possible. The general goal is to find a combination of weights that will result in the smallest error, by looking for a minimum point in a multidimensional error surface. This task can be done with a back-propagation algorithm, which uses a gradient descent approach to identify a minimum on the error surface. An additional parameter of the model fitting is the weight decay that penalizes large weights and thus large input to neurons. Note that the extreme case of a regression neural network with no hidden layers will be equivalent to a ridge regression. A description of the category of neural network described here can be found in chapter five of (Ripley 1996). I chose to focus on a simple form of neural network called "single hidden layer feed-forward perceptron" in which the system has three layers, with only one hidden layer, as shown in Fig. 2.1. Feed-forward means that the nodes can be numbered so that all connections go one way from a lower node to one with a higher number. I chose to use a linear activation function. Although this algorithm has the capacity to handle highly nonlinear systems, withmy choice of activation function, the system will only be linear. The model was implemented using the R libraries "nnet" (Venables and Ripley 2002) and "e1071" (Dimitriadou et al. 2011). The model was optimized for the number of neurons in the hidden layer and the weight decay parameter. The number of iterations to fit each neural network was set to 200. In every case (for non-cross-validated and cross-validated predictions), the model was run 10 times and the computed GEBV averaged,

to take into account dependency on initial weight parameters. This model will be referred to as neural network (NNET)

When it was readily possible to include covariates in the models, namely for the ridge regression methods, for BayesCπ, and for wBSR, the same analyses described above were performed with and without a covariate accounting for the population structure. The calculation of the covariate used is described below.

Prediction Accuracy and Cross-Validation

The predictive ability of the models was assessed using the Pearson correlation coefficient between the observations and the cross-validated GEBVs and will be referred to as the accuracy. Some publications define accuracy as the correlation between the GEBVs and observed phenotypic values divided by the square root of the heritability (Dekkers 2007; Lorenzana and Bernardo 2009). Using such a definition of accuracy introduces an additional error due to the heritability computation. In addition, the adjustment would be identical across all methods and would not contribute to differentiating them.

To compute the accuracy, I used a 10-fold cross-validation. Each phenotypic dataset was randomly divided into 10 equal parts. Then the GEBVs for each fold were predicted by training the model on the nine remaining folds. The accuracy was computed in one step on the whole vector of predicted values. To take into account the identified population structure of my datasets, a stratified sampling was used in each of the identified subpopulations to ensure that each fold was representative of the entire dataset composition. For the diallels, I treated the different crosses as subpopulations. In the other cases, I used the R package mclust (Fraley and Raftery

2002) to identify subpopulations by hierarchical clustering using a parameterized Gaussian mixture models. The Bayesian information criterion was used to identify the optimal number of subpopulations as well as the optimal clustering model to use.

To ensure an accurate comparison of models, the same cross-validation folds were used for each model. The non-cross-validated correlation was calculated as the correlation between the GEBVs obtained by using the whole dataset as a training population and the observed values on the training population. The difference between this non-cross-validated correlation and the accuracy was used as a measure of overfitting. For each dataset–trait combination I compared the GEBV estimates and the marker effect estimates between models. Marker effect estimates used in this comparison were computed for each model and dataset–trait combination using the whole dataset, as this would be the standard procedure in a GS application (use of all the data available to train the model). Significance of the differences in accuracies obtained with different models was tested using a binomial test for each pair of models using the accuracies obtained with each of the 18 traits as observations and considering the sign of the difference between accuracies; the null hypothesis is that the difference is not significant and then that the sign of the difference follows a Bernoulli distribution with parameter 0.5.

Accuracy within Subpopulations

The predictive ability of each of the tested GS models was also considered at the level of each subpopulation or cross. The homogeneity of variance among subpopulations was assessed with a Fligner-Killeen test (Conover et al. 1981) that is robust to nonnormality of the data. In addition, a significance test for the difference in accuracy among subpopulations was based on a randomization method as follows. The null

hypothesis is that the genetic parameters came from a single statistical population. Define $Var(X)$ as the variance of the accuracy measured across subpopulations. For each randomization k, randomly assign individuals to a population and calculate $Var(X_k)$. The probability of observing $Var(X)$ by chance alone is

$$p-value = \frac{1+number\ \ of\ \ randomization\ \ for\ \ which\ \ Var(X_k) \geq Var(X)}{1+n}$$

with n number of randomizations (Manly 1991). I used 10,000 permutations to compute the *p*-values.

The relevance of subpopulation structure was investigated with the pairwise F statistics quantifying the differences in allele frequency among subpopulations (F_{st}), estimated with a jackknifed estimator, as well as with a test of significance of the subpopulation structure on the differentiation using the R package hierfstat (Goudet 2005) with 1000 permutations. For all hypotheses testing, the Bonferroni correction for multiple testing was used.

Model Similarity

The similarities between models were also investigated through the use of clustering methods. For each of the 18 dataset–trait combinations, a matrix of Euclidean distances between GS models was calculated based on the cross-validated GEBVs. The GEBVs for each model were standardized to zero mean and unit variance before distance computation. Those distance matrices were then averaged (equal contribution of each dataset–trait combination) and used as an input for hierarchical clustering using the Ward criterion (i.e., based on the increase of variance of the cluster being merged during the tree building process). The tree built by equal contribution of each dataset (as opposed to dataset–trait combination) gave the same tree topologies. The

similarities were also analyzed by considering separately the traits whose genetic architecture was known to be characterized by some major effect loci, such as plant height in wheat, flowering time in the biparental *Arabidopsis thaliana* (L.) Heynh. population, and the betaglucan content in barley (*Hordeum vulgare* L.). The average excess kurtosis of the marker effect distributions obtained from ridge regression, wBSR, the BL, and BayesCπ as described below was also used to confirm this separation.

The similarities identified between models were further investigated by analysis of the marker effect distribution for each model, using the excess kurtosis that is a measure of the "peakedness" of the distribution. (A normal distribution has an excess kurtosis of 0.) Higher kurtosis means that more of the variance is the result of few extremely deviant marker effect estimates. I hypothesized that for some models, high kurtosis could be linked with the high multicolinearity of the data. This hypothesis was tested using a nonparametric correlation test based on Spearman's rho between the observed kurtosis and the number of lines, number of markers, and a statistic measuring the number of uncorrelated variables in the model. To construct this latter statistic, I used the number of eigenvectors necessary to capture 95% of the variance on a principal component analysis (PCA) of the marker dataset as an indicator of the number of uncorrelated variables.

Model Combinations

Considering the large diversity of GS models, instead of identifying a single best performing model, it could be advisable to build predictors based on a combination of models to increase the prediction accuracy. Various procedures to combine models were tested using the cross-validated GEBVs. To avoid introducing overfitting in the combined predictor, in all cases it was constructed using the same cross-validation

folds used to train the individual models, by building a combined predictor for each fold using the nine remaining folds as a training set. This procedure allowed us to make a direct comparison between the accuracies of the individual models and of the combined predictors. I tested the simple averaging of two to four models. I also used a simple least squares approach by regressing the cross-validated predictors on the observed data on each of the training sets. To obtain a more parsimonious model, I also used a backward stepwise model selection with the Aikake information criterion starting from the complete regression model described above. The R package MASS (Venables and Ripley 2002) was used to carry out this procedure. The approach of stacked regression described by (Breiman 1996a) was also used to build a combined predictor. This method is similar to the least square method described above, but with the regression coefficients constrained to be positive to account for the high colinearity between the predictors. (Breiman 1996a) reported a decrease of 10% in the prediction error with this approach. This modified least square approach was implemented using the optim function in R with a box constraint (Byrd et al. 1994) on the sign of the regression coefficients. Finally, in the light of the difference in prediction accuracies across subpopulations, a modified stacked regression method was tested giving equal weight to each subpopulation in the least square equation instead of giving an equal weight to each individual.

In addition to combining different models I sought to improve the accuracy of single models using a technique known as bagging in the machine learning literature (Breiman 1996b). This approach is the basis of the RF algorithm. For a given GS model it consists of generating training sets from the original dataset by sampling with replacement, with the size of the training set being equal to the original training set. The bagged predictor is then constructed by averaging the predictors obtained on the

different training sets. Breiman (1996b) showed that bagging effectively improved prediction accuracy of an unstable learning algorithm where a small perturbation in the training set can cause significant changes in the predictions. In an attempt to increase accuracy at the subpopulation level, I tested both a uniform sampling on the dataset and an equal sampling at the subpopulation level to obtain more balanced training sets.

Finally, I tried an approach called boosting (Drucker 1997), or AdaBoost in the machine learning literature, that was reported to be at least equivalent and in most cases superior to bagging in reducing the prediction error. In this approach, the model is trained repeatedly on the same sample. After each iteration, a measure of prediction error is computed for each individual. In the following iteration, the individuals with the highest error are given more weight in the training of the model. Over iterations, patterns that are more difficult to predict are given more importance and different machines are better in different parts of the observation space. The different predictors are combined using the weighted median such that those predictors with a reduced error are given more importance. This technique has been initially developed for classification purposes but has received an extension to regression problems. There are different versions of the boosting algorithm for regression; in this study I used AdaBoost.R2 (Drucker, 1997). Even though this algorithm is not reported to be the best one, it has the advantage of not requiring the set up of additional parameters relative to the error function used to update the weights given to each genotype in the training population (Shrestha and Solomatine 2006). A detailed presentation of the algorithm used can be found in Shrestha and Solomatine (2006). The original paper proposed several functions to compute the error, linear, squared, and exponential. All

three functions were tested. One potential drawback from this approach is a sensitivity to noise and outliers as the reweighting is proportional to the prediction error.

To test both bagging and boosting, I used the BL with the same parameters as described above for the BL alone and in the same cross-validation setting as for the other models. Thus, for each fold, the nine remaining folds were used as a training set. As for the single models, I also computed the non-cross-validated correlation as an additional measure of overfitting.

All statistical procedures were executed using R (R Development Core Team, 2010). The executable for wBSR was obtained from the authors Hayashi and Iwata (2010).

Results

Ridge Regression Models

The accuracies of the three different ridge regression methods were quite similar. Across the traits, the average of the accuracies was 0.55 for the BRR and 0.56 for ridge regression with grid search and the ridge regression using a mixed model to estimate the shrinkage factor. Across all traits, the median of the correlations between the cross-validated GEBVs obtained using the three different ridge regression methods was above 0.96. The comparison of marker effects for those ridge regression approaches also demonstrated their high similarities: the correlation between marker effects was above 0.99 for all traits except for betaglucan (correlation of 0.78). The kurtosis of the marker effect distribution was in the same range for each method with an average excess kurtosis of 1.41 for the grid search case and 1.55 for the two other methods. This means that the marker distribution was on average slightly less

"peaked" for the grid search version of the ridge regression. In addition, the non-cross-validated correlations were similar between those methods for all traits. The computation time was considerably lower for the ridge regression using a mixed model: the grid search ridge regression took approximately half of the computation time of the Bayesian ridge and the ridge regression with a mixed model took only one third of the time required to do the ridge regression with a grid search.

BayesB and Weighted Bayesian Shrinkage Regression

The results of the grid search to optimize the prior parameters of wBSR revealed a wide range of accuracies from 0.48 for the average accuracy across traits with the best performing common combinations of prior parameters (π = 0.25, ν = 9, S^2 = 0.05) to an average of 0.56 for the best traitwise performing combinations. The π parameter (prior probability that a marker will have a null effect) and the scale parameter of the prior are the most important parameters according to the complete grid search made on the different traits (600 sets of prior parameters tested). The best values of the scale parameters were between 0.001 and 10 with most of them close to the value of 0.043 used by Meuwissen et al. (2001). The best π value ranged from 0.01 to 0.99. Figure 2.2 presents heat maps made by averaging across traits the accuracies obtained with the grid search. Since no best set of parameters used a scale parameter greater than 0.1, the heat maps exclude all grid points with a scale parameter above 0.1. The black dots in Fig. 2.2 indicate the best set of prior parameters identified for each trait. It is interesting to notice that those dots are relatively scattered across the heat maps, visualizing the fact that no single parameter setting was best for all traits.

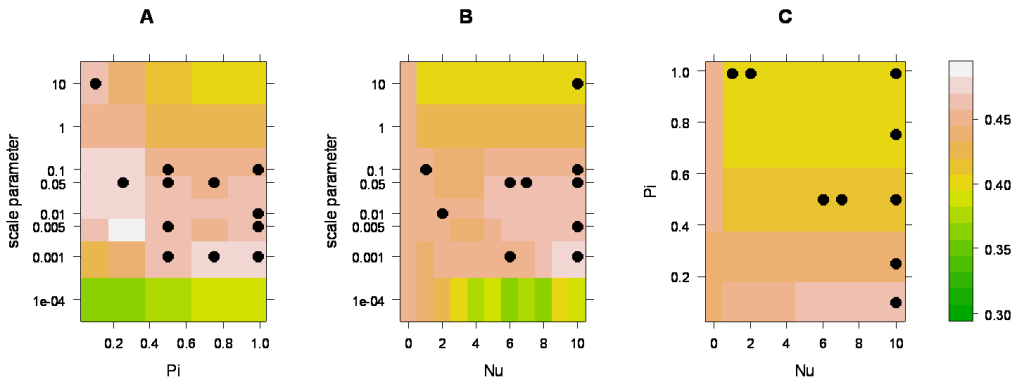


Figure 2.2. Heat maps summarizing accuracies for the grid search on weighted Bayesian shrinkage regression (wBSR) prior parameters. The accuracies were centered across the 600 triplets of parameters for each trait and averaged. The axes of the heat map correspond to the prior parameters for the marker effect variance. Nu, the degree of freedom, and the scale parameter are the parameters of the scaled inverse chi square distribution of the marker effect variance prior and Pi the prior proportion of loci with a null variance. Each cell of the heat maps is an average across the nonplotted parameter values. The scale parameter is plotted on a log scale (the set of prior parameter of BayesB is Pi = 0.9, Nu = 4.36, S2 = 0.01). The black dots indicate for each trait the best set of prior parameters identified.

Overall, and on average, the set of prior parameters of BayesB (Pi = 0.9, Nu = 4.36, S2 = 0.01) was approximately a good value. The average accuracy for the best set of prior parameters for each trait was 0.56 compared to 0.45 using the original prior parameter set of BayesB in wBSR. It was not computationally feasible to compute BayesB for the Wheat CIMMYT, Barley CAP, and Wheat Cornell datasets. For other datasets, the average accuracy for BayesB was 0.52 compared to 0.53 for wBSR with the prior parameter set of BayesB and 0.71 for the best set of prior parameters for each trait, as identified by the grid search. The difference between the non-cross-validated correlation and the accuracy, taken as a measure of overfitting (a lower value indicates less overfitting), was also favorable to wBSR: 0.08 for the best set of prior parameters

for each trait for wBSR, 0.12 for wBSR with the prior parameter set of BayesB, and 0.18 for BayesB itself.

The sensitivity of wBSR to the marker order was tested for all datasets by randomizing the marker order in the design matrix. In all cases, the correlation between cross-validated GEBVs using different marker orders was above 0.98. For some datasets, however (Wheat diallel [Limagrain Europe] and Wheat Cornell), the correlation between marker effects themselves was only around 0.6. Averaging of marker effects from several runs with different marker order did not improve wBSR accuracy. This result suggested that the algorithm always captures the same signal, but if several markers are in high linkage disequilibrium (LD) with a QTL, the algorithm tended to pick the first marker entered in the model. Given that the focus of GS is on the GEBVs, this is probably not an issue for the use of this algorithm for GS purposes. Similar findings were reported in the use of VBay, which is an EM algorithm equivalent to BayesCπ developed for genome-wide association studies approaches in humans (Logsdon et al. 2010). This finding prevented further direct comparison of marker effects between wBSR and other models. Distributions of marker effects could be compared, however. The average excess kurtosis of the marker effects distribution for BayesB was 38.2 compared to 19.42 for wBSR with BayesB prior and 8.76 for wBSR with an optimized prior. The correlation between GEBVs from BayesB and wBSR with the prior parameter set of BayesB was high and ranged from 0.77 to 0.95 except for the moisture trait in the Panel maize dataset where it was only 0.62.

Empirical Bayes Grid Search

The results of the grid search to optimize the prior parameter of E-Bayes revealed a wide range of accuracies from 0.46 for the average accuracy across traits with the best performing common combinations of prior parameters (τ = –0.5, ω = –0.5) to an

average of 0.54 for the trait wise best performing combinations. Figure 3.3 presents a heat map made by averaging across traits the accuracies obtained with the grid search. On average, the set of original set of prior parameters (i.e., τ = 2 and ω = 2 in the bottom left of the heat map) was close to the optimum, with an average accuracy of 0.46. This set of prior parameters corresponds to a flat (noninformative) prior.

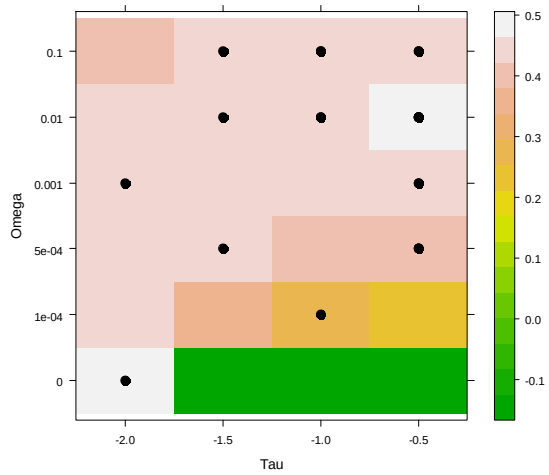


Figure 2.3. Heat map summary of the grid search on empirical Bayes (E-Bayes) prior parameters. The accuracies were centered across the 24 pairs of parameters for each trait and averaged. The axes of the heat map correspond to the prior parameters for the marker effect variance. τ is the degree of freedom and ω is the scale parameter of the marker effect variance prior. Each cell of the heat maps is an average across the nonplotted parameter values. (The original set of prior parameter is τ = –2, ω = 0). The omega parameter is plotted on a log scale. The black dots indicate for each trait the best set of prior parameters identified.

Comparison of Accuracies and Overfitting

Table 2.2 presents the accuracy obtained for each trait and model tested. For the sake of clarity and considering the similarity of the three ridge regression models, only the ridge regression using a mixed model to estimate the shrinkage parameter is given

here. Hereafter, "wBSR" and "E-Bayes" denote the optimized versions, with a different set of prior parameters for each trait. The last three lines of the table give the average accuracy, the average non-cross-validated correlation (providing a measure of the overfitting level for each model), and the average MSE.

Most models reached a very similar accuracy for a given trait. However, RKHS tended to outperform the other models in terms of accuracy. Support vector machine performed poorly on these datasets, even though the model was optimized for each trait for the cost and epsilon parameter SVM was the only method significantly different from all the other for the accuracy ($p < 0.05$) with Bonferronni correction for multiple testing. The others pairs of methods significantly different from each other ($p < 0.05$), with Bonferronni correction for multiple testing, were wBSR from the elastic net, RKHS from E-Bayes, elastic net, and the neural network. The performance of the elastic net model was slightly below that of ridge regression and the BL. The relative percentage of lasso penalty ranged from 0.7 to 1 (pure lasso) with an average of 0.92 across traits. Meanwhile, if I consider the difference between the non-cross-validated correlation and the accuracy as a measure of overfitting, E-Bayes, RKHS, SVM, and NNET are clearly overfitting much more than the other models.

Table 2.2. Accuracy for each trait and model, average non-cross-validated correlation for each model, and average MSE for each model.

‡Betaglucan, betaglucan content; FLOSD, flowering time in short days; DM10, dry matter in nonlimiting N conditions; DM3, dry matter in limiting N conditions; YLD1 to YLD5 refers to the yield traits reported in Crossa et al. (2010); TKW, thousand kernel weight.

§RR-BLUP, random regression best linear unbiased predictor; BL, Bayesian Lasso; wBSR, weighted Bayesian shrinkage regression; E-Bayes, empirical Bayes; RKHS, reproducing kernel Hilbert space; SVM, support vector machine; RF, random forest; NNET, neural network.

Dataset†	Trait‡	RR-BLUP§	BL	Elastic net	wBSR	BayesCπ	E-Bayes	RKHS	SVM	RF	NNET
Barley 1	Yield	0.53	0.6	0.52	0.53	0.53	0.53	0.6	0.43	0.6	0.51
Barley CAP	Betaglucan	0.57	0.6	0.57	0.57	0.57	0.57	0.6	0.35	0.6	0.54
Bay × Sha (Bay-0 × Shahdara)	FLOSD	0.82	0.8	0.83	0.83	0.82	0.82	0.83	0.8	0.9	0.82
	DM10	0.63	0.6	0.63	0.64	0.63	0.63	0.64	0.56	0.6	0.56
	DM3	0.4	0.4	0.4	0.4	0.39	0.4	0.41	0.33	0.4	0.35
Panel maize	Moisture	0.75	0.8	0.75	0.76	0.75	0.73	0.79	0.45	0.7	0.73
	Yield	0.63	0.6	0.61	0.63	0.63	0.59	0.64	0.32	0.6	0.59
Diallel maize	Moisture	0.74	0.7	0.72	0.73	0.74	0.73	0.75	0.56	0.6	0.72
	Yield	0.52	0.5	0.49	0.51	0.52	0.51	0.5	0.29	0.5	0.48
Wheat CIMMYT	YLD1	0.51	0.5	0.46	0.48	0.51	0.49	0.59	0.36	0.5	0.54
	YLD2	0.5	0.5	0.45	0.5	0.5	0.46	0.52	0.36	0.4	0.51
	YLD4	0.38	0.4	0.35	0.36	0.38	0.36	0.43	0.32	0.4	0.43
	YLD5	0.44	0.5	0.42	0.47	0.44	0.39	0.52	0.27	0.5	0.44
Wheat Cornell	Yield	0.36	0.4	0.37	0.37	0.34	0.26	0.28	0.22	0.4	0.36
	Height	0.45	0.4	0.41	0.44	0.44	0.41	0.55	0.37	0.5	0.45
Wheat diallel	Height	0.64	0.7	0.68	0.67	0.66	0.67	0.73	0.51	0.6	0.67
	TKW	0.6	0.6	0.59	0.6	0.59	0.59	0.68	0.41	0.5	0.65
	Yield	0.53	0.5	0.51	0.52	0.53	0.51	0.58	0.39	0.5	0.57
Average accuracy (cross-validated)		0.56	0.6	0.54	0.56	0.55	0.54	0.59	0.41	0.5	0.55
Average non-cross-validated correlation		0.77	0.8	0.75	0.77	0.77	0.93	0.99	0.89	0.8	0.85
Average MSE		0.67	0.7	0.69	0.68	0.68	0.76	0.64	1.36	0.7	10.54

The MSE was computed on the scaled phenotypic data and cross-validated GEBVs centered and scaled by the phenotypic variance. This scaling ensured that the traits with a higher phenotypic variance were not weighted more heavily. Most models were rather similar but SVM and NNET performed poorly in terms of MSE. As the data were centered and scaled with the phenotypic variance before MSE computation, it did not measure the bias between the cross-validated GEBVs and the phenotypic data but only the error and the difference in the level of shrinkage between models. As the average accuracy of NNET is similar to the best performing model, I can attribute the higher MSE on average to a higher variance of the cross-validated GEBVs, even when scaled by the phenotypic variance.

Comparison of Cross-Validated Genomic Estimated Breeding Values

The comparison of cross-validated GEBVs between models allowed estimation of the similarities and dissimilarities of the models. To determine whether model similarities were affected by genetic architecture, I analyzed separately traits that were believed to be influenced by major loci versus traits that were unlikely to be affected by major loci as described in the material and methods. The dendrogram topologies were extremely similar for these two categories, indicating that, at least at the crude level explored here, genetic architecture did not affect model similarity.

Figure 4.4 presents the hierarchical clustering tree obtained through the averaging of the distance matrix across all traits. Those results clearly showed the similarities in terms of GEBVs between the linear models represented by ridge regression and the hierarchical Bayesian methods and the distinctness of nonparametric methods such as RF, neural network, and RKHS regression. Note too that while the nonparametric methods cluster with each other, they are all quite different, with deep divisions in the

clustering between each method. Support vector machine was not used in this analysis as its poor prediction performance would have clustering difficult to interpret. This analysis also showed the strong similarity between ridge regression and BayesCπ. The similarities between wBSR and the BL is also interesting as wBSR was grid searched for the optimal prior parameters but the BL was not. It is also interesting to note that the elastic net clustered with E-Bayes despite being a combination of lasso and ridge regression penalty.

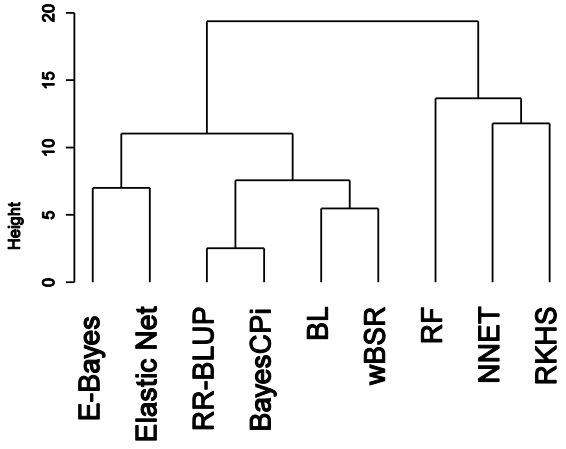


Figure 2.4. Hierarchical clustering of genomic selection (GS) models based on cross-validated genomic estimated breeding values (GEBVs), the height on the y axis refers to the value of the criterion associated with a particular agglomeration of models. RR-BLUP, random regression best linear unbiased predictor; BL, Bayesian Lasso; wBSR, weighted Bayesian shrinkage regression; RF, random forest; NNET, neural network; RKHS, reproducing kernel Hilbert space.

Comparison of Marker Effect Distributions

The distribution of the excess kurtosis of marker effects for each GS model that estimates marker effects was studied across the 18 dataset–trait combinations. The ridge regression marker effect distribution rarely departs from a normal distribution excess kurtosis (0), whereas other models such as the BL, wBSR and to some extent BayesCπ displayed significant differences in the marker effect distribution according to the trait. This suggests that Bayesian learning was taking place for the BL, wBSR, and to some extent for BayesCπ. Empirical Bayes and the elastic net performed differently, which is consistent with the clustering results and were characterized by an extremely high kurtosis. This is consistent with the variable selector properties of the elastic net.

The relationships between the excess kurtosis of different model marker effect distributions were investigated (Fig. 5.5). It is important to note that although the linear correlation was presented here, the excess kurtosis was not a linear function of the marker distribution. This figure revealed a striking behavior of E-Bayes and elastic net compared to the other models. I expected a significant correlation across trait–dataset combinations between kurtosis of different models. High and significant correlations were observed for ridge regression, BayesCπ BL, and wBSR but not for E-Bayes or the elastic net. As E-Bayes seemed to be characterized by more overfitting than the other models, I investigated the impact of the number of lines, number of markers, and a measure of the number of uncorrelated variables in the marker dataset on the number of PCA axes needed to capture 95% of the variance. For all models except E-Bayes and elastic net, the Spearman correlation between these variables and the excess kurtosis was not significant (p-values > 0.4). For E-Bayes the correlations were significant, with p-values of 0.011, 0.04, and 0.009 for the number of lines, the

number of markers, and the number of uncorrelated variables, respectively. A multiple regression with these variables captured 38% of the variance of the E-Bayes marker effect distribution excess kurtosis. These correlations constitute evidence that E-Bayes did not handle highly multidimensional data well and tended to capture more noise than the other models. For the elastic net, the correlations were also significant with p-values of 0.0001, 0.09, and 0.02 for the number of lines, the number of markers, and the number of uncorrelated variables, respectively. A multiple regression with these variables captured 72% of the variance of the elastic net marker effect distribution excess kurtosis. This can be related to the formulation of the lasso that can retain only as many variables as observations. However, the elastic net performed correctly in terms of accuracy whereas E-Bayes did not.

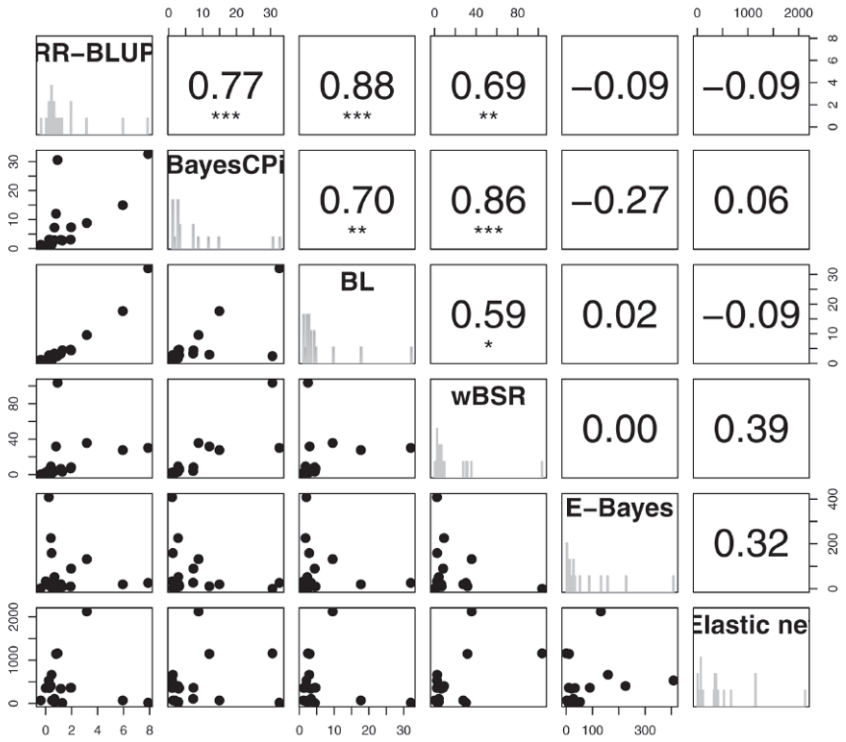


Figure 2.5. Comparison of marker effect distribution. For each model, histogram of the marker effect distribution excess kurtosis on the diagonal, scatter plots comparing two models below the

diagonal (each point represents one trait–dataset combination), and Spearman correlation between models above the diagonal with the significance level of the correlation based on Spearman's rho. (*Significant at the 0.05 probability level; **Significant at the 0.01 probability level; ***Significant at the 0.001 probability level. RR-BLUP, random regression best linear unbiased predictor; BL, Bayesian Lasso; wBSR, weighted Bayesian shrinkage regression.

Prediction in Each Subpopulation

My clustering approach revealed no genetic structure in the Bay × Sha and Panel maize datasets.

Genomic selection accuracy was strongly affected by subpopulation (Table 3.3). This observation was true across all GS models: all models tended to be better for some subpopulations than for others with some trait–dataset combinations showing extremely high differences in accuracy. For example, in the Wheat Cornell yield dataset I distinguished six subpopulations, one of which had an accuracy of 0.7 and two that had an accuracy of 0.0. For the other traits and datasets, the differences in accuracy were less striking but in most cases the accuracy varied by at least twofold between the best and worst predicted subpopulations, even for the diallel design. Across all models and traits the standard deviation of accuracy between subpopulations ranged from 0.05 to 0.3.

Table 2.3. Summary of the subpopulation results. The accuracies reported here are from the Bayesian Lasso.
†Barley CAP (Barley Coordinated Agricultural Project, 2011); Wheat CIMMYT (Crossa et al., 2010); Wheat Cornell (Heffner et al., 2011); Wheat diallel, Limagrain Europe, Chappes, France.
‡Betaglucan, betaglucan content; YLD1 to YLD5 refers to the yield traits reported in Crossa et al. (2010), Bonferroni correction for multiple testing; TKW, thousand kernel weight.

Dataset†	Trait‡	Number of groups	Smallest group size	Biggest group size	Minimum accuracy	Maximum accuracy	Total accuracy	*p*-value of subpopulation effect on accuracy	Fligner-Killeen test for phenotypic variance homogeneity
Barley CAP	Betaglucan	6	61	285	0.44	0.64	0.57	1	0.69
Wheat CIMMYT	YLD1	7	38	161	0.32	0.53	0.5	1	1
	YLD2				0.4	0.59	0.49	1	1
	YLD4				0.26	0.44	0.37	1	1
	YLD5				0.29	0.57	0.47	1	1
Wheat Cornell	Yield	6	38	95	–0.11	0.7	0.35	1.08×10^{-2}	0.08
	Height				0.21	0.45	0.44	1	2.73×10^{-5}
Wheat diallel	Height	8	32	82	0.11	0.69	0.66	2.40×10^{-3}	2.82×10^{-14}
	TKW				0.13	0.62	0.6	3.84×10^{-2}	0.49
	Yield				0.01	0.55	0.52	2.40×10^{-2}	1.09×10^{-2}
Diallel maize	Moisture	5	48	114	0.38	0.8	0.74	1.20×10^{-3}	0.53
	Yield				0.18	0.63	0.51	0.34	1

I considered three hypotheses to explain the subpopulation effect on GS accuracy. First, subpopulations with higher phenotypic variance might also have higher genetic variance and strongly influence the models, which would in turn lead them to have higher accuracy. Second, subpopulations that were larger would have more individuals in the training population and would have higher accuracy. Third, subpopulations with higher average pairwise Fst values would be more genetically unrelated to the population as a whole and therefore have lower accuracy. I tested the three hypotheses simultaneously using multiple regression of subpopulation accuracy on the three variables within each dataset–trait combination but no variable showed a consistently

significant or suggestive relationship with accuracy using a Bonferroni correction for multiple testing. However, the two traits that displayed a significant difference in accuracy despite a nonsignificant heterogeneity of variance, the moisture content trait for the Diallel maize dataset and the thousand kernel weight trait for the Wheat diallel dataset, had significant p-values for the effect of the Fst before Bonferroni correction (0.06 and 0.08 respectively).

Use of Structure Covariate

The fraction of phenotypic variance explained by the structure covariates I used was below 5% for all traits considered and models except for Barley CAP with wBSR, in which it reached 14%, and the Diallel maize with BayesCπ, in which it reached 18 and 16% for moisture content and grain yield, respectively. The use of a structure covariate did not improve the accuracy inside each subpopulation compared to a model without a covariate (data not shown). This result suggested that the GS models are able to capture the subpopulation structure information most of the time and that this information can contribute to a significant portion of the accuracy.

Combinations of Models

The various differences observed between GS models suggest that complementarities exist between them that could be used to improve accuracy. Nevertheless, in most cases, combining different models did not result in a gain in accuracy. The only gains observed were for flowering time in Bay × Sha where the accuracy went from 0.85 with the best single method to 0.9 with a combination and the Wheat Cornell yield dataset where the accuracy went from 0.36 to 0.39. The accuracy at the subpopulation level was not improved, even with the modified version of the stacked regression estimator designed to favor the accuracy at the subpopulation level.

Bagging and Boosting

The results of bagging applied to the BL did not bring any additional accuracy gain. The average accuracy reached with the BL alone was 0.56, while with bagging the average accuracy dropped to 0.28 and the non-cross-validated correlation increased from 0.78 to 0.84. However, important differences in the impact of bagging were observed between the structured and nonstructured datasets. Bagging minimally reduced accuracy in nonstructured datasets (to 0.49 from 0.56) whereas in structured datasets bagging reduced the average accuracy to 0.01. The use of an alternative bagging strategy to bootstrap samples equally in each of the subpopulations, however, did not bring any improvement of the accuracy.

The boosting of the BL did not bring an improvement of the accuracy for either the median or the average approach used to combine the different predictors or for any of the loss functions tried. The average accuracy was below 0.1 for the linear and exponential loss functions and was equal to 0.2 for the squared loss function. Here again, the nonstructured datasets were better predicted than the structured ones. For none of the traits considered did bagging or boosting bring an increase in accuracy.

Discussion

One of the key results of this study was that despite similar average accuracies between most of the models tested, there were major differences between them in terms of cross-validated GEBVs and marker effects.

Choice of a Genomic Selection Model for Plant Breeding

An optimal GS method should provide the highest accuracy possible, limit overfitting on the training dataset, and be based as much as possible on marker-QTL LD rather than on kinship (Habier et al., 2007). Moreover, such methods must be easy to implement, reliable across a wide range of traits and datasets, and computationally efficient. To be implemented, it should be possible to run the models overnight for the datasets I used. Model sparsity has been advocated as a key criterion for method selection. Sparsity can be achieved in two ways. First, it could arise from the elimination of markers with small effects from the model. This way is not favorable because, for polygenic traits, small or partial-effect markers do explain some true genetic variance not captured by large-effect markers. Second, it could arise from the capacity of the method to ensure that markers in strong LD with large QTL can capture their full effect rather than allowing the effect to be distributed over a number of markers. This way is favorable and might be observed by measuring excess kurtosis. These two ways are not mutually exclusive.

These general guidelines would lead to the recommendation to use RR-BLUP with a mixed model, the BL for its versatility, and wBSR. The elastic net performed well and it produced extremely sparse models, much more so than its Bayesian counterparts. From a breeding point of view, a nonsparse model could be more favorable: with more

markers being selected by the model, more time will be required to reach fixation. In addition, (Legarra et al. 2010) stressed that conditional expectations are optimal for selection(Gianola and Fernando 1986). Conditional expectations of the GEBVs based on the observations maximize expected selection response based on truncation selection via maximization of the correlation between predictor and predictand. These can be obtained through the BL but not with the regular Lasso or elastic net.

The use of BayesCπ cannot be recommended considering the extremely high similarities with RR-BLUP and the increased computation time. The high overfitting observed with E-Bayes as well as the observation that excess kurtosis was driven by marker colinearity suggested that this model should not be used in its current form. The high overfitting observed with the neural network approach would suggest that this model should not be used for GS at this time. In addition, the high computing requirement, mainly because of the model optimization step for training the neural network, also precluded its use for GS.

The case of RKHS regression is more difficult because, even though the model was overfitting, its accuracy was higher, indicating that the model was capturing both more genetic signal and more noise than the other models. This problem might be addressed by the use of different kernels and distance functions. An advantage of RKHS regression is that it is performed on the individual rather than the marker space. Thus, it is feasible to implement RKHS regression even if the dataset segregates for millions of markers, as would be the case in species where LD decays rapidly.

Despite overall good results and a reasonable computing time, the RF should be used with caution considering that this is a new method for GS. However, the apparent distinctness of this method and its potential to capture nonadditive effects, compared to the more classical approaches, should encourage more development.

Kurtosis of Marker Effects

The variation of the excess kurtosis is of importance as it signals if a model was able to adjust the marker effect distribution to the distribution of the QTL effects. For a given level of accuracy, it seems reasonable to favor a model whose marker effects are closer to the distribution of the QTL effects. A variable excess kurtosis is also an indirect indicator of the basis of the accuracy of a given model. If excess kurtosis varies with the traits, it suggests that an important part of the accuracy is based on LD marker-QTL association rather than on kinship.

Need for Further Analysis on the Basis of Accuracy for the Best Models Selected

These results would need to be confirmed using an approach similar to Habier et al. (2007) to identify the basis of predictive ability of models. If the predictive ability of a given model is based mainly on kinship, it will decrease much faster than if the predictive ability of the model is based on LD between markers and QTL. In addition, if a model is based on kinship rather than marker-QTL LD, the increase in inbreeding due to the application of a GS scheme using such model will be much faster. The simulation results of Habier et al. (2007) suggested that the accuracy of Bayesian methods such as BayesB (Meuwissen et al., 2001) would be based more on marker-QTL LD than on kinship, while that of ridge regression (RR-BLUP) is based mainly on kinship. The results of simulations (Long et al. 2011) suggest that most of the BL accuracy is due to LD marker-QTL.

Population Substructure

The large difference in accuracy observed in some cases between subpopulations in this study deserves additional analysis to uncover the basis for those differences. This cannot be explained by an uneven sampling of the cross-validation folds because the

sampling approaches I used ensured that each fold was representative of the total dataset composition. All models were similar in terms of their differences in accuracy among the subpopulations. However, on a trait-by-trait basis, not all models performed the same in the subpopulations. The small number of dataset–trait combinations considered precluded broader conclusions on this observation. Clearly more investigation is needed to uncover the basis for those differences in accuracy. I observed that subpopulation accuracy differences were trait dependent for the same marker dataset (e.g., yield versus height for the Wheat Cornell dataset). Furthermore, the distribution of the polymorphism information content values, minor allele frequencies, and marker and individual call rates were roughly similar in the subpopulations. Together, these observations suggest that subpopulation accuracy differences could be caused by differences in the genetic determination of the trait in each subpopulation as well as by differences in phenotypic variance. For the Wheat Cornell dataset, the F_{st} values between the well predicted subpopulations and the poorly predicted subpopulations were somewhat higher than the other pairwise F_{st} values, suggesting that they were more differentiated, but such elevated F_{st} values were not found in other traits (data not shown). Excluding the poorly predicted subpopulations from the training set did not affect the cross-validated accuracy in the remaining part of the dataset (data not shown). In addition, using only the best predicted subpopulation did not result in accuracy as high as with the complete dataset for those populations. Thus, even data from subpopulations that are poorly predicted contribute beneficially to prediction accuracies. With the data available, it was not possible to clearly distinguish what part of the gain in accuracy was associated with the increase in the training population size and what part was due to the use of a more diverse training dataset with more recombination events than in any single subpopulation.

For only two traits (moisture content for the Diallel maize dataset and thousand kernel weight for Wheat diallel dataset), the difference in phenotypic variance between subpopulations was not significant while the difference in accuracy was significant. For both traits, the multiple regression approach only allowed us to suggest a negative correlation between F_{st} and the accuracies.

Overall, these observations suggest that the difference in accuracy cannot be explained only by a difference in phenotypic variance in subpopulations but rather by difference in genetic architecture between subpopulations. This difference in genetic architecture is also more likely to exist when two given subpopulations are more unrelated than others, which would account for the observed F_{st} pattern. The nonsignificance of the multiple regression approaches precluded drawing a strong conclusion on the origin of those differences in accuracy.

Given that I do not understand the basis for differences among subpopulation accuracies, it seems necessary when implementing a GS model to focus not only on the overall accuracy but also on the accuracy at the level of each subpopulation as an additional check of model accuracy. This result has potentially wide ranging implications. For example, differences in accuracy could affect the rate of inbreeding generated by GS. If only some of the subpopulations of a breeding program are predicted well by a GS model, it may lead to preferential selection from those subpopulations and to the loss of diversity represented by the other subpopulations as most of the candidates lines identified by the GS model and confirmed in the field will originate from those well predicted subpopulations. I do not argue that differences in accuracy among subpopulations should preclude GS in structured populations and restrict it to biparental populations: loss of genetic diversity is also a risk in biparental GS and could even be greater because of the already reduced genetic diversity within any given cross.

Combination of Models

I was disappointed by the efficiency of the combination of predictor approaches that I tested. The lack of gain in accuracy is interesting in and of itself as it suggests that all models tested capture the same signal but in different ways, as shown by the differences between GEBVs. As discussed above, the differences in the capture of the signal, for example through kinship or marker-QTL LD, have important implications. This is additional evidence supporting the use of a few models based more on marker-QTL LD than on kinship as both will capture the same signal but in ways that may have different consequences for successful breeding.

As my study was unable to identify an all-purpose model or combination of models, I would suggest that for implementation in breeding programs the BL or wBSR with a grid search should be used. Results from RF seem promising but need more study with simulated datasets to better understand the genetic basis of the accuracy with this model (kinship or marker QTL LD). As this model does not produce marker effects it was not possible to investigate that point by studying the variation of the excess kurtosis of the marker effects distribution across traits. In addition, RKHS could potentially capture nonadditive relationships but the predictions obtained would not be GEBVs.

Bagging and Boosting

The lack of gain of accuracy by the use of bagging and boosting is also interesting as an indication that by a simple use of the BL model I already reach a plateau in terms of accuracy. For the boosting, this indicates that the poorly predicted individuals do not carry any additional genetic signal that can be effectively captured by the BL. Breiman (1996b) and Drucker (1997) acknowledge that neither bagging nor boosting

can transform a poor predictor into a good one in all cases. In addition, those approaches were mainly developed for so called "weak learners," that is, predictors that are only weakly correlated with the true value. I are not sure that this definition applies to the BL. However, this definition is quite arbitrary. In addition, recent work on various boosting algorithms applied to classification (Long and Servedio 2009) demonstrated that convex potential boosting algorithms such as AdaBoost are sensitive to noise in real datasets. Thus, my study can only conclude that bagging and boosting of the BL are not useful for GS. However, this approach could be useful to enhance the predictive ability of simpler models as reported by (González-Recio et al. 2010).

Acknowledgments

The authors thank P. Flament, S. Chauvet, and all the Limagrain Europe biostatistics team for their helpful suggestions. The USDA-NIFA-AFRI provided grant support (award numbers 2009-65300-05661 and 2011-68002-30029). Additional funding for this research was provided by USDA-NIFA National Research Initiative CAP grant No. 2005-05130 and by Hatch 149-402. Part of this work was carried out by using the resources of the Computational Biology Service Unit at Cornell University, which is partially funded by Microsoft Corporation.

References

Barley Coordinated Agriculture Project. 2011. Introduction to project. Available at http://www.barleycap.org (verified 26 Oct. 2013). Univ. of Minnesota, St. Paul, MN.

Breiman L (2001) Random forests. Mach Learn doi: 10.1023/A:1010933404324

Breiman L (1996a) Stacked regressions. Mach Learn doi: 10.1023/A:1018046112532

Breiman L (1996b) Bagging predictors. Mach Learn doi: 10.1007/BF00058655

Bureau A, Dupuis J, Falls K, et al. (2005) Identifying SNPs predictive of phenotype using random forests. Genet Epidemiol doi: 10.1002/gepi.20041

Byrd RH, Lu P, Nocedal J, Zhu C (1994) A mimited-memory algorithm for bound constrained optimization. SIAM J Sci Comput doi:10.1137/0916069

De los Campos G, Naya H, Gianola D, et al. (2009) Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics doi: 10.1534/genetics.109.101501

Conover WJ, Johnson ME, Johnson MM (1981) A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. Technometrics doi:10.2307/1268225

Crossa J, de Los Campos G, Pérez P, et al. (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics doi:10.1534/genetics.110.118521

Dekkers JCM (2007) Prediction of response to marker-assisted and genomic selection using selection index theory. J Anim Breed Genet doi: 10.1111/j.1439-0388.2007.00701.x

Dimitriadou E, Hornik K, Leisch F, et al. (2011) e1071: Misc functions of the department of statistics (e1071), TU Wien. R package version 1.6. Available at http://CRAN.R-project.org/package=e1071 (verified 26 Oct. 2013). R Foundation for Statistical Computing, Vienna, Austria.

Drucker H (1997) Improving regressors using boosting techniques. In: Fisher Jr. DH (ed) Proc. 14th Int. Conf. Mach. Learn. (pp. 107. Morgan Kaufmann, San Mateo, CA, pp 107–115

Drucker H, Burges CJC, Kaufman L, et al. (1997) Support vector regression machines. Adv Neural Inf Process Syst 9:155–161.

Fraley C, Raftery AE (2002) Model-based clustering, discriminant analysis, and density estimation. J Am Stat Assoc doi: 10.1198/016214502760047131

Friedman JH, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. J Stat Softw 33:1–22.

Gardner M., Dorling S. (1998) Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmos Environ doi: 10.1016/S1352-2310(97)00447-0

Gianola D, de Los Campos G, Hill WG, et al. (2009) Additive genetic variability and the Bayesian alphabet. Genetics doi: 10.1534/genetics.109.103952

Gianola D, Fernando RL (1986) Bayesian methods in animal breeding theory. J Anim Sci 63:217–244.

González-Recio O, Forni S (2011) Genome-wide prediction of discrete traits using bayesian regressions and machine learning. Genet Sel Evol doi: 10.1186/1297-9686-43-7

González-Recio O, Weigel K, Gianola D, et al. (2010) L2-Boosting algorithm applied to high-dimensional problems in genomic selection. Genet Res doi: 10.1017/S0016672310000261

Goudet J (2005) a package for R to compute and test hierarchical F -statistics. Mol Ecol Notes doi: 10.1111/j.1471-8278 .2004.00828.x

Habier D, Fernando RL, Dekkers JCM (2007) The impact of genetic relationship information on genome-assisted breeding values. Genetics doi: 10.1534/genetics.107.081190

Hayashi T, Iwata H (2010) EM algorithm for Bayesian estimation of genomic breeding values. BMC Genet doi:10.1186/1471-2156-11-3

Heffner EL, Jannink J-L, Sorrells ME (2011) Genomic selection accuracy using multifamily prediction models in a wheat breeding program. Plant Gen doi: 10.3835/plantgenome2010.12.0029

Heffner EL, Sorrells ME, Jannink J-L (2009) Genomic selection for crop improvement. Crop Sci doi: 10.2135/cropsci2008.08.0512

Hornik K (1989) Multilayer feedforward networks are universal approximators. Neural Netw doi: 10.1016/0893-6080(89)90020-8

Institut National de la Recherche Agronomique (INRA). 2007. Web Service VNAT. Study of the natural variation of Arabidopsis thaliana. Available at http://dbsgap.versailles.inra.fr/vnat/ (verified 24 Oct. 2013). INRA, Paris, France.

Jannink J-L, Lorenz a. J, Iwata H (2010) Genomic selection in plant breeding: from theory to practice. Brief Funct Genomic Proteomic doi: 10.1093/bfgp/elq001

Kang HM, Zaitlen NA, Wade CM, et al. (2008) Efficient control of population structure in model organism association mapping. Genetics doi: 10.1534/genetics.107.080101

Legarra a., Robert-Granié C, Croiseau P, et al. (2010) Improved Lasso for genomic selection. Genet Res doi: 10.1017/S0016672310000534

Liaw A, Wiener M (2002) Classification and regression by RandomForest. R News 2:18–22.

Logsdon B a, Hoffman GE, Mezey JG (2010) A variational Bayes algorithm for fast and accurate multiple locus genome-wide association analysis. BMC Bioinform doi: 10.1186/1471-2105-11-58

Long N, Gianola D, Rosa GJM, Weigel K (2011) Long-term impacts of genome-enabled selection. J Appl Genet. doi: 10.1007/s13353-011-0053-1

Long PM, Servedio R a. (2009) Random classification noise defeats all convex potential boosters. Mach Learn doi: 10.1007/s10994-009-5165-z

Lorenz A. J, Chao S, Asoro FG, et al. (2011) Genomic selection in plant breeding : knowledge and prospects. Adv Agron doi: 10.1016/B978-0-12-385531-2.00002-5

Lorenz a. J, Hamblin MT, Jannink J-L (2010) Performance of single nucleotide polymorphisms versus haplotypes for genome-wide association analysis in barley. PLoS One 5:e14079. doi: 10.1371/journal.pone.0014079

Lorenzana RE, Bernardo R (2009) Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theor Appl Genet doi: 10.1007/s00122-009-1166-3

Loudet O, Chaillou S, Camilleri C, et al. (2002) Bay-0 x Shahdara recombinant inbred line population: a powerful tool for the genetic dissection of complex traits in Arabidopsis. Theor Appl Genet doi: 10.1007/s00122-001-0825-9

Manly BFJ (1991) Randomization and Monte Carlo methods in biology. Chapman and Hall/CRC, London, UK

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–29.

Moser G, Tier B, Crump RE, et al. (2009) A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. Genet Sel Evol doi: 10.1186/1297-9686-41-56

Park T, Casella G (2008) The Bayesian Lasso. J Am Stat Assoc doi: 10.1198/016214508000000337

Pérez P, de los Campos G, Crossa J, Gianola D (2010) Genomic-enabled prediction based on molecular markers and pedigree using the Bayesian linear regression package in R. Plant Gen J doi: 10.3835/plantgenome2010.04.0005

Plummer M, Best N, Cowles K, Vines K (2006) CODA: Convergence diagnosis and output analysis for MCMC. R News 6:7–11.

R Development Core Team (2010) R: A language and environment for statistical computing. Available at http://www.r-project.org (verified 18 Oct. 2013).R Foundation for Statistical Computing, Vienna, Austria.

Ripley BD (1996) Pattern recognition and neural networks. Cambridge Univ. Press, Cambridge, UK.

Shrestha DL, Solomatine DP (2006) Experiments with AdaBoost.RT, an improved boosting scheme for regression. Neural Comput doi: 10.1162/neco.2006.18.7.1678

Smola AJ, Schölkopf B (2004) A tutorial on support vector regression. Stat Comput doi: 10.1023/B:STCO.0000035301.49549.88

Venables WN, Ripley BD (2002) Modern applied statistics with S, Fourth edi. Springer, New York, NY

Xu S (2007) An empirical Bayes method for estimating epistatic effects of quantitative trait loci. Biometrics doi: 10.1111/j.1541-0420.2006.00711.x

Xu S, Hu Z (2010) Methods of plant breeding in the genome era. Genet Res (Camb) doi: 10.1017/S0016672310000583

Yi N, Xu S (2008) Bayesian LASSO for quantitative trait loci mapping. Genetics doi: 10.1534/genetics.107.085589

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Stat Soc Ser B doi: 10.1111/j.1467-9868.2005.00503.x

CHAPTER 3

USING GENOMIC PREDICTION TO CHARACTERIZE ENVIRONMENTS AND OPTIMIZE PREDICTION ACCURACY IN APPLIED BREEDING DATA[3]

Abstract

Simulation and empirical studies of genomic selection (GS) show accuracies sufficient to generate rapid annual genetic gains. Whole-genome genotyping provides the opportunity to go beyond the evaluation of lines to the evaluation of alleles and thus, provides new tools to analyze multi-environment trials (MET). Considering allele replication rather than line replication provides a new way to cope with highly unbalanced phenotypic datasets. Using a two-row elite barley (*Hordeum vulgare* L.) population representative of the type of data generated by a commercial breeding program and tested for grain yield across Europe from 2007 to 2010, we characterized allele effect estimates at each test location and used them to successfully identify outlier environments. I also used the prediction accuracy between environments to characterize the environments. The prediction accuracy gave the same pattern as the genetic correlation between environments based on a factor analytic model, suggesting that it could be used to cluster environments. A new method was developed to optimize the composition of the training population for predicting performance in the target population of environments (TPE). This method does not search for mega-environments, but instead it identifies and removes less predictive environments from the set of environments used to train the model. Using this approach with the barley dataset, cross-validated accuracy increased from 0.54 to 0.61 while controlling overfitting and focusing the prediction on the TPE. This study demonstrates the possibilities offered by GS to analyze MET, identify outliers, group environments, and select historical data relevant for current breeding efforts

[3]Heslot N, Jannink J-L, Sorrells ME (2013) Using genomic prediction to characterize environments and optimize prediction accuracy in applied breeding data. Crop Sci doi: 10.2135/cropsci2012.07.0420

Abbreviations

AIC, Akaike criterion; BL, Bayesian lasso; BLUE, best linear unbiased estimator; DH, doubled haploid; GEBV, genomic estimated breeding value; G*E, genotype by environment interactions; GS, genomic selection; MET, multi-environment trials; QTL, quantitative trait locus; SNP, single nucleotide polymorphism; TBV, true breeding value; TPE, target population of environments.

Introduction

The genomic selection (GS) concept was first proposed by (Meuwissen et al. 2001). The basis of this approach is to estimate breeding values for quantitative traits based on whole genome genotypes through the simultaneous estimation of many marker effects. Simulations and empirical studies have demonstrated that GS could greatly accelerate the breeding cycle, maintain genetic diversity within the breeding programs, and increase genetic gain beyond what is possible with phenotypic selection or quantitative trait loci (QTL) mapping approaches (Heffner et al. 2009; Lorenz et al. 2011)

With GS, it becomes possible to take advantage of the large amount of phenotypic data collected by breeding programs across years, provided a source of DNA or genotypic data is available. However, this also raises new challenges to optimally exploit those data. Historical data will include trials of varying quality without readily accessible meta-information on issues affecting trial quality. By nature, historical data are also extremely unbalanced. In addition, some data may not be relevant for the target population of environments (TPE) (Comstock 1977). Consequently, the sampling of environments in large datasets might not reflect their expected frequency in the current breeding target, potentially skewing selection pressures. In the context of dramatically decreasing genotyping costs, phenotyping is becoming the most

expensive step in plant breeding and being able to extract the most information from phenotypic data is becoming more crucial.

One of the key issues in plant breeding is genotype by environment interactions (G*E), i.e., the frequent observation in large MET that genotypes react unequally to dissimilar environments, leading to differences in scale between environments and rank changes among genotypes (Cooper and DeLacy 1994). G*E are called non-cross-over when no rank changes occur (scale differences) and cross-over when rank changes occur. The latter especially complicates selection for broad adaptation, because the mean performance of a genotype across environments might not select the best performing genotype everywhere.

A very large number of methods have been developed to study and cope with G*E (Cooper and Hammer 1996; van Eeuwijk et al. 2005). The most recent development is the use of the mixed model framework to analyze G*E, particularly multiplicative mixed models, such as the factor analytic model, to account for the covariance (or lack thereof) between environments responsible for G*E (Beeck et al. 2010; Burgueño et al. 2008; Kelly et al. 2009; Piepho 1998). A recent paper (Burgueño et al. 2012) reported on the convergence of this approach with the GS framework to improve GS prediction accuracy, by using a relationship matrix based on markers in the mixed model.

Obviously, GS will not change the fact of G*E; it could, however, provide new tools to analyze datasets affected by G*E and therefore enable better selection decisions in its presence. To test approaches that use GS for this purpose, we used a two-row elite barley population from a commercial breeding program. In addition to multiple years of field evaluation, each line was genotyped with a moderate number of markers. The number of markers we had available was far fewer than might be obtained from, for example, next-generation sequencing (Elshire et al. 2011). Nevertheless, published

results on breeding populations that have small effective population sizes suggest that high levels of accuracy can be attained even with a reduced set of markers (Heslot et al. 2012; Lorenz et al. 2012). In addition, the simultaneous estimation of marker effects on the whole genome is more central to the GS concept than the specific number of markers.

Using this data, the objectives of this study were to 1) propose and evaluate a method for analyzing large unbalanced multi-environment trials in the GS context, based on allele effect variation across environments in order to identify outlier environments, 2) compare the pattern of G*E observed with GS with that observed directly on phenotypes, and 3) optimize the use of these data for genomic selection by focusing the prediction on the performance in the TPE. This study was carried out using a large two-row spring barley dataset as a case study example. This dataset is typical of the data generated by a commercial breeding program in that it is unbalanced and composed of advanced breeding lines and commercial check varieties.

Materials and methods

Phenotypic and genotypic data

A dataset consisting of grain yield data of 996 F6, F7, and doubled haploid (DH)-derived elite two-row spring barley breeding lines grown in 58 different European environments (combinations of years and locations) from 2007 to 2010 (total of 11,570 adjusted means) was provided by Limagrain Europe (Chappes, France). The trials had been managed using standard growing practices, including fungicide treatments. The experimental design in each environment was an alpha-lattice. Adjusted means were computed for each location, taking into account the experimental design and were used as raw data in this study. Because the dataset was

derived from early generation yield trials performed across several years, it was extremely unbalanced, with only 18 lines out of the 996 present in more than half of the environments. Each trial data passed Limagrain's standard minimum quality control for discarding failed trials and extreme outliers (A.-M. Bochard, personal communication, 2012). This quality control standard is lax and it was assumed that some outlier trials remained. To validate results from analyzing the 2007 to 2010 datasets, we used an independent, unbalanced dataset of 212 lines grown in 16 European locations in 2011. These validation trials used the same experimental design as did the trials from which modeling data were obtained.

The lines were genotyped with 335 single nucleotide polymorphisms (SNPs), providing whole genome coverage. For each marker, missing data were imputed as the mean of the non-missing data. Although other imputation procedures may be slightly more accurate, mean imputation is standard in GS work and was adequate for this dataset because of the small numbers of missing data (3% missing on average per genotype and 8% missing on average per marker).

Phenotypic analysis of the dataset

To provide a clear framework to the new approaches we propose, a mixed model framework was used to analyze the data using ASReml (Gilmour et al. 2009), and variance component estimates based on all lines and all environments were extracted. Each year-location combination was defined as a single environment.

The following model (Model 1) was used:

$Y = \mu 1_n + Zs + W_1 u + W_2 \gamma + I_n \varepsilon$,

with μ being the overall mean, n the number of observations, s the environment effect, g the number of lines, j the number of environments, Z the environment design matrix that is equal to $I_j \otimes 1_g$ in the balanced case, with $\otimes$ Kronecker product,

u the line effect with design matrix W_1 with one column for each line, which is equal to $1_j \otimes I_g$ in the balanced case, and has variance σ_g^2, γ the effect of the G*E interaction, with design matrix W_2, which is an identity matrix if the data are fully balanced. Here it has n rows and $s*g$ columns. It can be constructed by considering the identity matrix with $s*g$ rows, and removing the rows corresponding to unobserved combinations. The effect has variance σ_{ge}^2, and ε the residual. This term accounts for environment-specific line effects. 1_n is a vector of 1 with n elements and I_n is the identity matrix with n rows.

For that model, the environmental effects were assumed to be independent and normally distributed with mean 0 and variance equal to σ_e^2. The line effects were sampled from a multivariate normal distribution with mean 0 and covariance proportional to the realized relationship matrix A based on molecular markers. A was calculated as the product of the marker design matrix, normalized by allele frequency with its transpose. To solve singularity issues in the realized relationship matrix, a small scalar (10^{-5}) was added to the diagonal elements using the procedure suggested in (Piepho et al. 2012). The use of the realized relationship matrix allows the computation of a G*E interaction effect despite the imbalance in the data. The covariance matrix for the G*E was $I_j \otimes A$, where j is the number of environments and $\otimes$ denotes the Kronecker product of matrices. Residuals are assumed i.i.d. and normally distributed with variance σ_ε^2.

The broad-sense heritability, H, was estimated according to (Hallauer et al. 2010) across the j environments, assuming no replications, as follows:

$H = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_{ge}^2/j + \sigma_\varepsilon^2/j}$ This equation assumes balanced data, but it is justified for unbalanced data if σ_ε^2/j is small relative to σ_g^2 and σ_{ge}^2/j (Holland et al., 2002)

The genetic variance-covariance matrix between locations was modeled in another model:

$Y = \mu 1_n + Zs + W_2 u + I_n \varepsilon$

The covariance of the genetic effects u was defined as $G_j \otimes A$. G_j was the variance covariance matrix for the environments. G_j was modeled using the identity matrix, i.e., without accounting for G*E (Model 2), or with a diagonal matrix to account for heterogenous variances between environments (Model 3). G_j was also modeled with a factor analytic model (Model 4) to study the genetic correlation between environments, and thus, the patterns of G*E (Beeck et al. 2010; Cullis et al. 2010) (Model 4). For the factor analytic modeling of order k, $G_j = \Lambda\Lambda' + \Psi$ where Λ is a j by k matrix containing the environment loadings for the kth factor and Ψ is a diagonal matrix with different non-negative parameters on the diagonal.

Different levels of model complexity were tested: with no modeling of the covariance, heterogenous variances between environments, or 1 or 2 factor analytic components. The best model was identified using the Akaike information criterion (AIC). The genetic correlation matrix between environments was obtained for Model 4 as described in (Cullis et al. 2010) using the R package myf. Because of computational and numerical challenges when fitting such a complex model on a very large and unbalanced dataset, the dataset was reduced to 61 lines that were present in more than a third of the environments for this analysis only. With the full dataset, even for the one component factor analytic model, it was not possible to get the model to converge

because of singularity issues. This was also the case even when providing informed starting values for the model, such as the variance and loadings obtained from the same model with a reduced dataset. This reduced set of lines was referred to as "checks" and represented 13.9% of the total number of entries in the complete dataset. With this reduced dataset, 55% of the cells were missing in the location-mean matrix of 61 lines by 58 environments.

Genomic selection model

Based on the results from Heslot et al. (2012), the Bayesian Lasso (BL) (Park and Casella 2008) was chosen as the genomic selection model, using the implementation provided in the R package "BLR" (Pérez et al. 2010). The model is of the form:

$$Y = \mu 1_n + X\beta + I_n\varepsilon$$

Inmyanalyses, Y was the mean phenotype of a line in an environment. For analyses of a single environment, μ is the environment mean. For analyses with more than one environment, best linear unbiased estimators (BLUEs) for the lines were computed using ASReml on the adjusted mean per location and used as Y . X was the marker design matrix, β was the vector of marker effects, and the error term, ε, was assumed to be normally distributed with mean 0 and variance equal to σ^2. The estimator of β is $\hat{\beta} = \arg\min_{\beta}\left(\left(\tilde{y} - X\beta\right)'\left(\tilde{y} - X\beta\right) + \lambda\left\|\beta\right\|_1\right)$ with $\tilde{y} = y - \bar{y}1_n$. The $\arg\min_{\beta}$ notation refers to the determination of coefficients β minimizing the expression inside the brackets. The shrinkage is marker specific and dependent on a regularization parameter. This regularization parameter is based on the L_1 norm (also named Taxicab or Manhattan norm): $\left\|\beta\right\|_1 = \sum_i \left|\beta_i\right|$. Relative to ridge regression, the BL produces weak shrinkage of regression coefficients with large absolute values and strong shrinkage of coefficients with values near zero, leading to a sparse model. In

the Bayesian version of the Lasso, each marker effect β_j is assigned a normal prior of mean 0 and variance σ_j^2. Each σ_j^2 follows an independent exponential prior with parameter $\lambda^2/2$ and λ is further assigned a Gamma prior. Estimates were obtained using a Gibbs sampler based on Markov Chain Monte Carlo (MCMC). The model was run for 20,000 iterations and the first 5,000 iterations were discarded as burn-in, and the chains were not thinned. Model convergence was visually assessed based on the trace of parameter samples across iterations

Prediction Accuracy and Cross-validation

The predictive abilities of the training populations were assessed using the Pearson correlation coefficient between the observations and the cross-validated genomic estimated breeding values (GEBV). This correlation will be referred to as the accuracy. In selection theory, the accuracy is defined as the correlation between the selection criterion and the true breeding value (TBV). As calculated here, this correlation is reduced by deviations of the phenotype from the TBV. In principle, an unbiased estimate of the accuracy can be obtained by dividing the correlation between GEBV and phenotype by the square root of the heritability (Dekkers, 2007; Lorenzana and Bernardo, 2009). The correlation was not adjusted in this way in this study because it introduces an additional error due to the heritability computation.

A 10-fold cross validation was used to compute the accuracy. Each training dataset (subset of environments or complete dataset) was randomly divided into 10 equal folds. Then, the GEBVs for each fold were predicted by training the model on the nine remaining folds. There was no clear subpopulation structure of individuals warranting a stratified sampling of individuals for cross-validation (data not shown). The correlation between prediction and phenotype was computed in one step on the whole vector of predicted values.

Use of marker effects to characterize environments

The BL was used to estimate marker effects in each of the environments separately. The differences and similarities between environments were characterized using a clustering approach based on the marker effects. Because complete marker data were used for all 996 lines, the marker effects in each environment formed a balanced dataset, enabling the computation of a Euclidean distance matrix between environments. By doing so, each marker was treated as a descriptor of the environment. Heat maps were produced to facilitate the interpretation of the results using the R package gplots (Warnes 2001). I also computed the prediction accuracy between pairs of environments. The prediction was based on the analysis of one environment while excluding from the correlation those lines common to both environments to compute the accuracy, as those lines were used in the training environment. This gave a reciprocal prediction accuracy matrix between environments.

Because entries differed across environments, we tested the effect of the genetic composition on the accuracies between environments as well as on the marker effect-derived Euclidean distance. A Mantel test was used to determine significance between similarity matrices and the matrices of pairwise F_{st} (Wright's F_{st} for population differentiation), pairwise G_{ij} (mean kinship), and pairwise D_s (standard genetic distance;(Nei 1978)) between locations. Those statistics were computed using SPAGeDI (Hardy and Vekemans 2002). The mean kinship G_{ij} is equivalent to averaging the kinship coefficients of the kinship matrix computed as the product between the marker design matrix and its transpose for the lines present in each pair of environments. Because the accuracies between environment matrices were not symmetrical, we constructed two symmetrical matrices using the prediction in one

direction or in the other and averaged them before performing the test. A similar test was also carried out between the accuracy matrix and the environment correlation matrix obtained from the mixed model analysis of the 61 checks dataset to determine whether both methods were capturing the same G*E pattern.

A method to optimize accuracy

For each environment, we calculated its predictive ability as the mean accuracy in predicting line performance in each of the other environments, and environments were ranked accordingly (Figure 3.1). This rank was used to separate the total dataset into predictive and unpredictive subsets as follows. Data from all environments were initially placed in the predictive subset. Then, starting with the least predictive environment, data were moved, one environment at a time, from the predictive to the unpredictive set. The cross-validated accuracy using the BL was computed on the predictive set, using a 10-fold cross-validation. The cross-validation folds were identical, as long as all the lines were retained in the predictive set. At each step, the model built on the predictive set was also used to predict a BLUE computed on the unpredictive set. If prediction accuracy on the unpredictive set increased, it indicated that some relevant information had been moved to the unpredictive set. Moving environments from predictive to unpredictive sets was terminated when cross-validated accuracy within the predictive set decreased and accuracy on the unpredictive set increased. Two approaches were evaluated, one with the complete initial dataset and another with prior removal of the outlier environments identified using the cluster analysis based on marker effects. A validation dataset, consisting of progeny lines phenotyped in 2011, was used to test whether this approach increased prediction accuracy for datasets outside of the training data, and to verify that overfitting did not occur. Here overfitting was defined as the loss of predictive power

outside of the training data as a result of the model capturing noise or signal relevant only to the training dataset.

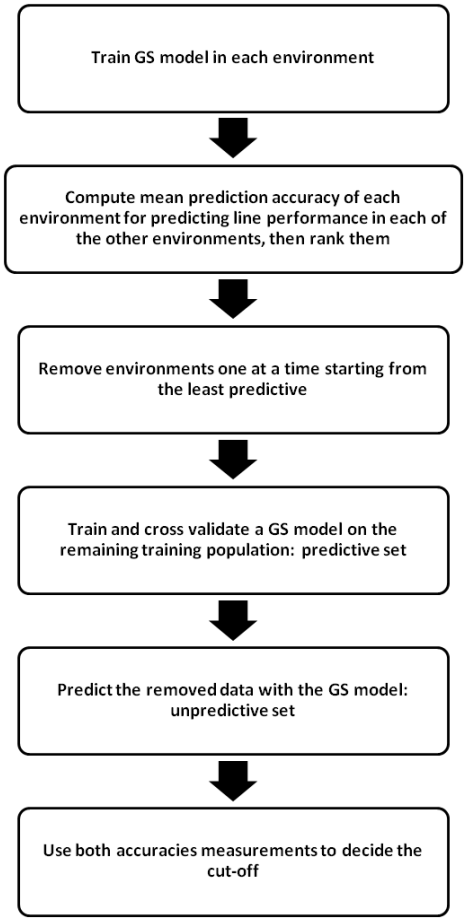


Figure 3.1. Representation of the optimization procedure used. GS, genomic selection.

As an additional validation of this procedure, 25 environments were selected at random from the complete dataset. Each environment, in turn, was removed prior to applying the procedure and then predicted using the newly identified optimal model. If the procedure did not improve prediction, we would expect a 50% probability that model optimization would increase or decrease (null hypothesis) accuracy in any given environment. This null hypothesis was tested using a binomial test that allowed

a comparison between the full data set and the optimal models across a broader range of environments than available from the 2011 validation dataset. All statistical procedures were executed using R (R Development Core Team 2010).

Results

Impact of G*E on accuracy

Analysis of all the data and the realized relationship matrix yielded the following variance component estimates using Model 1: the additive genetic variance was 2.4, the environment variance 7.3, the additive G*E variance was 4.8 and the residual variance 20.3. On average across lines, this gives an entry-mean heritability of 0.84, associated with a G*E variance twice the genetic variance.

Using the reduced dataset of 61 lines, the multiplicative mixed model (Model 4) that was a best fit based on the AIC was a model with a first order factor analytic modeling of the covariance between environments. For the base model with no G*E term and no covariance modeling (Model 2), the AIC was 7024.6 going up to 7060 for the diagonal variance model (Model 3) but down to 6895.1 with a factor analytic 1 (FA1) model (Model 4). It was not possible to correctly fit a higher-order factor analytic model, even by providing informed starting values for the parameters, such as the variance components from the FA1 model or by fixing some variances. In addition, the factor analytic regression of order one accounted for more than 99% of the genetic variance in 48% of the environments, using the approach described in Beeck et al. (2010) This result further suggested that the optimal model was of first order and explained the difficulty in fitting a more complex model.

I identified a striking lack of a relationship between the mean prediction accuracy of the environments (using one environment for training) and the training population size (number of lines in the training environment) (Figure 3.2). The overall accuracy level was low (maximum around 0.2), but the training populations were small (50 to 400 individuals). It is important to note that the matrices of prediction accuracy between

environments with or without common lines were very strongly correlated (p < 10-16).

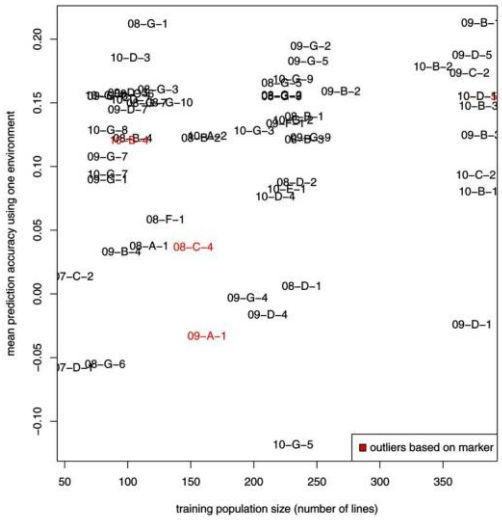


Figure 3.2. Relationship between the mean prediction accuracy of the environments, using one environment for training and the training population size (number of lines in the training environment). The first number for each environment refers to the year, followed by a letter for the geographic area, and ending with a number for the trial site. The environments in red are those identified as outliers using marker effects clustering.

Characterization of environments using genomic selection approaches

Figure 3.3 represents a heat map of pairwise environment distances, computed using marker effects. The analysis clearly differentiated four environments (combination of years and locations) from the rest of the dataset and suggested several large groups of environments. However, no geographic, crop management or weather pattern explained the large groups of environments. The meta-data available included breeders' notes on each of the environments. The environments isolated from the rest of the data by marker-effect clustering were confirmed as outliers by breeders for different reasons, including poor trial establishment and heavy rains. There were no outliers identified by the breeders that were not identified by the marker effects.

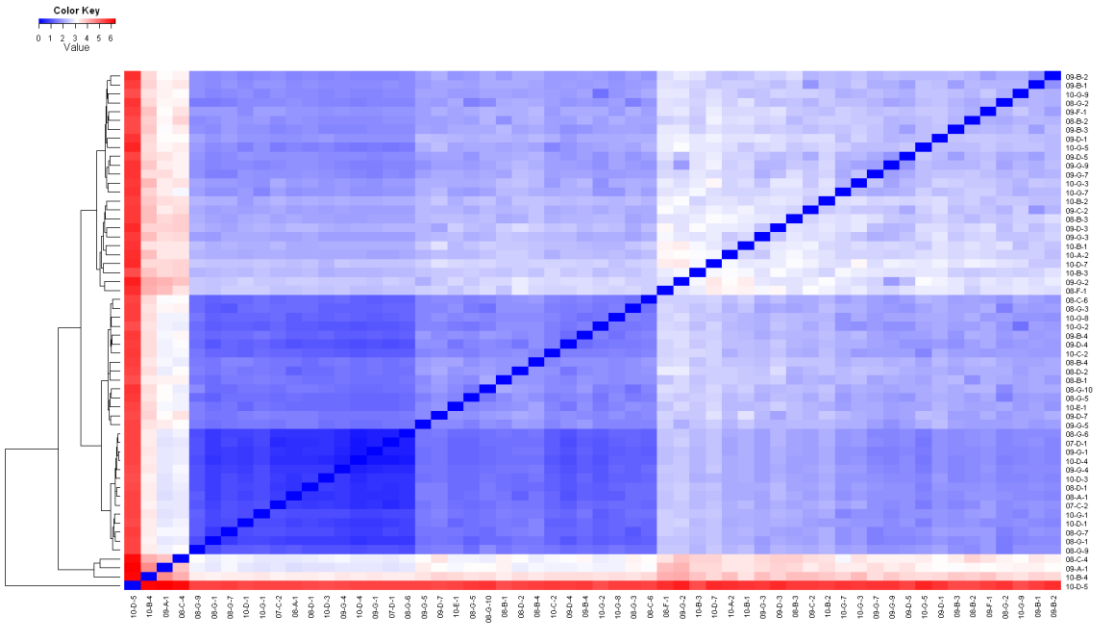


Figure 3.3. Heat map of environments based on Euclidean distances computed using marker effects. The environment codes are formed by concatenating the two last digits of the year, a letter for the geographic area, and a number for the trial site. Red indicates larger distance, whereas blue indicates smaller distance between environments.

A heat map of the prediction accuracy between pairs of environments when the common lines between the training and the predicted environments were removed from the predicted environment is shown in Fig. 3.4. The upper and lower triangles of the heat map represent the prediction accuracies in each direction. In this case, the pattern is not as clear as in the heat map of Euclidean distances between environments (Figure 3.3). In both cases, a complete linkage clustering method was used looking for compact and similar clusters. The outliers identified by the breeders and by the marker effects were the same. A Mantel test performed between the upper and lower triangle of the matrix rejected the null hypothesis that there is no correlation between the two matrices tested (p < 10-16). Therefore, prediction accuracies in both directions (predicting from environment A to B or from B to A) were significantly correlated. In addition, a Mantel test performed between the prediction accuracy matrix and the between-environment genetic correlation matrix obtained on the reduced phenotypic

dataset gave a *p*-value of $3*10^{-4}$ (correlation of 0.33). This correlation is low but highly significant, thus it is evidence that both approaches captured a similar pattern in the data, but not only that pattern, and that accuracy between pairs of environments was affected by the genetic covariance between those environments (but not only by it). I can conclude that for this dataset the pattern of G*E was the same for GS and phenotypic selection.

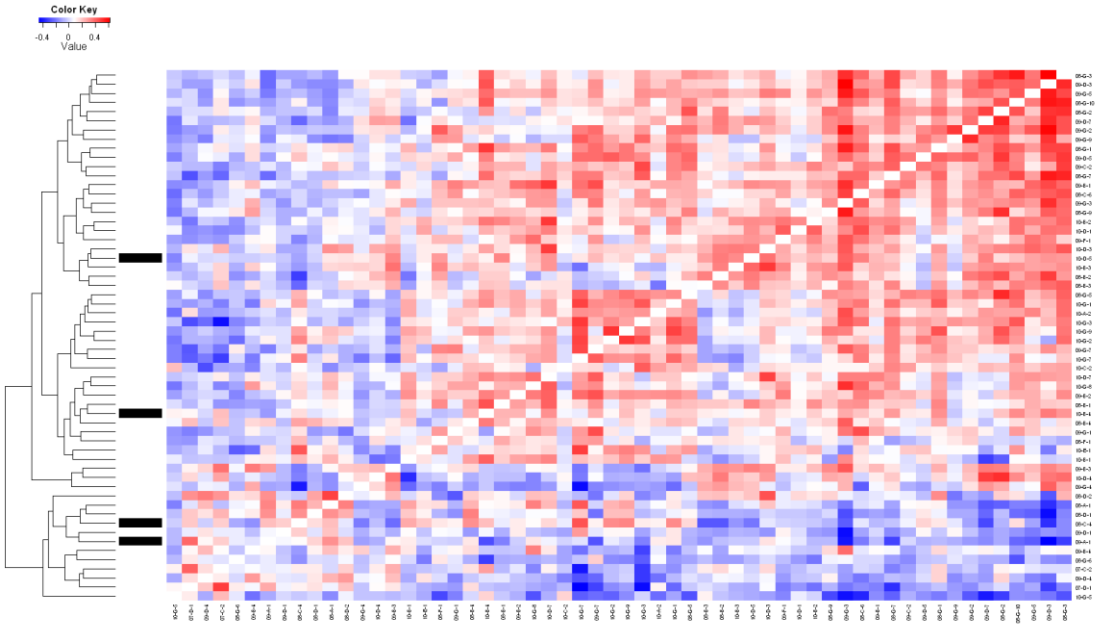


Figure 3.4. Heat map of the prediction accuracy between pairs of environments excluding common lines. The environment codes are formed by concatenating the two last digits of the year, a letter for the geographic area, and a number for the trial site. The black bars on the left indicate the environments identified as outliers by the clustering based on marker effects.

To examine the question of whether the genetic composition of lines tested in the environments might affect their prediction accuracy, different measures of relationships between environments were compared with their marker-effect distances and reciprocal prediction accuracies (Table 3.1).

Table 3.1 *P*-value of Mantel tests (100 000 permutations) between the environments' Euclidean distance matrices based on marker effects, the accuracy matrices between environments (with no lines in common), and different measures of genetic distance between environments: pairwise F_{st} between locations, pairwise G_{ij} (mean kinship) and D_s, the (Nei, 1978) standard genetic distance. The null hypothesis was that there is no correlation between the two matrices tested. The correlations are shown in parenthesis.

Differentiation measure	Marker effect distance	Reciprocal prediction accuracy
F_{st}	0.51 (0.09)	1 (-0.23)
G_{ij}	0.08 (-0.35)	0.009 (0.58)
D_s (Nei 1978)	0.50 (-0.30)	1 (0.55)

There was no correlation between the differences in the genetic composition of a given environment as measured by the pairwise F_{st} between locations or the pairwise standard genetic distance D_s (Nei, 1978) and the differences in marker effects between environments or the accuracy. There was a weakly significant correlation (p = 0.08) between the Euclidean distance based on marker effects and the mean kinship between locations (pairwise G_{ij}) (Table 3.1). The reciprocal prediction accuracy was impacted in part by the mean kinship between locations as well (p < 10-2).

Optimizing accuracy one environment at a time

As described in the Materials and Methods, this strategy involved calculating the mean prediction accuracy for each environment when predicting line performance in each of the other environments followed by ranking these predictive abilities. Cross-validated accuracy within the predictive dataset initially increased slowly with the number of environments removed, whereas accuracy of the predictive model on the unpredictive dataset initially increased, then reached a plateau at about 0.15, suggesting that very little useful information was removed from the predictive set (Figure 3.5). The cross-validated accuracy with all 58 environments obtained initially was 0.54, which

increased to 0.61 when 19 environments were removed. Even though 19 environments were removed from the training dataset, only one of the 996 barley lines was excluded from the training dataset. The broad-sense heritability of the optimal training population was 0.83 compared with 0.84 for the complete dataset. I refer to the model using the predictive set as the optimal model and contrast it to the full model that uses all data available. The prediction accuracy of the new lines evaluated across 16 locations in 2011 in the same geographic area went up from 0.28 with the full model to 0.29 with the optimal one, calculated as the correlation between the predicted values and a BLUE computed across the 16 locations of 2011. This difference in accuracy was not significant. I believe the empirical nature of the proposed training population optimization makes it difficult to validate. The slight increase in prediction accuracy, however, supported the fact that the method did not lead to overfitting, which would have caused the accuracy to decline. It is also worth noting that the year 2011 was characterized by an unusual pattern of drought in the spring and heavy rain in early summer. Thus, the 2011 set of validation environments might, in fact, have a low frequency in the TPE. Using the same strategy, but after prior removal of the outliers identified with the marker-effects approach previously discussed, did not provide any improvement. Note that the optimal predictive set actually contained some of the environments identified as outliers using marker effects.

As an additional validation criterion, the same optimization procedure was carried out 25 times, each time leaving one out of 25 randomly selected environments. An optimized model was then computed and its prediction accuracy was compared with the prediction accuracy obtained using the model trained on all the data minus that environment. A binomial test was used to determine whether the accuracy obtained using the optimal model was higher than when using the full data. That test gave a *p*-value of 0.014, rejecting the null hypothesis that the procedure had no impact on

accuracy. This is additional evidence that the proposed procedure provided a small but real gain in accuracy and that this approach is useful for optimizing the training population to increase prediction accuracy for performance in the TPE.

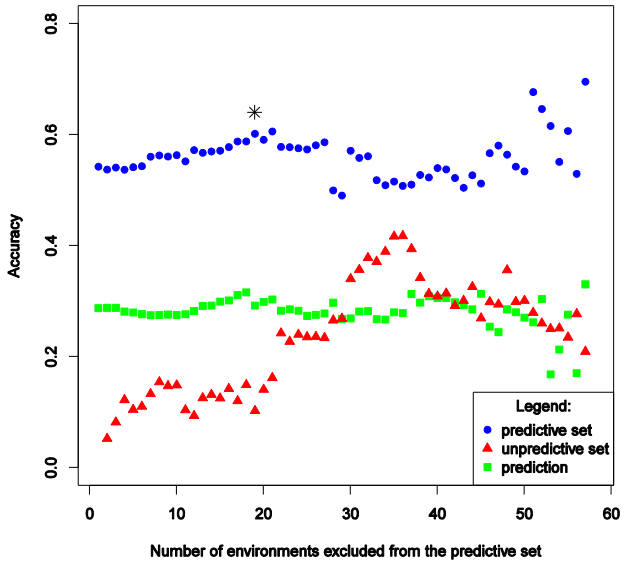


Figure 3.5. Results of the training population optimization approach: red dots are cross-validated accuracies in the remaining training population (predictive set); blue triangles are prediction accuracies for the data excluded from the training population (unpredictive set). Green squares are the prediction accuracies for the validation set (progeny lines observed in 2011).

The correlation between the full model and the optimal model for cross-validated GEBVs was 0.78 and that for marker effects was 0.74. The excess kurtosis of the marker-effects distribution dropped from 3.79 with the full model to 2.74 with the optimal model. A larger excess kurtosis meant that there were larger marker effects. Larger marker effects might be expected in response to unusual stresses that are not in the TPE. Similarly, large marker effects will have a disproportionate impact on clustering based on marker effects. This suggests some connection between the marker effects clustering and the approach based on accuracy.

Discussion

Evaluation of the importance of G*E in this dataset

The variance components provided evidence that the interaction between genotypes and environments was large in this dataset. The fact that the best multiplicative mixed model to account for environment covariance was of the first order suggests that the interaction was complex and was due, in large part, to a lack of correlation between environments (Cullis et al., 2010). This is the most problematic form of G*E interaction as it results in a change in rank for genotypes across environments.

To make the among-environment correlation matrix estimable, the dataset was restricted to the 61 lines occurring most frequently across environments. It is thus possible that the correlation matrix obtained was not representative of the G*E pattern of the complete dataset. Indeed, the lines retained were mostly commercial checks and cultivars about to be released. Such genotypes might produce a different pattern or a reduced amount of G*E compared to lines still under evaluation. Nevertheless, the among-environment correlation matrix we obtained was highly significantly correlated with the accuracy matrix between environments, suggesting that the G*E pattern of those lines was a good representation of the overall G*E pattern. It also suggests that the G*E patterns for phenotypic and genomic selection were similar and could be dealt with in the same way.

The impact of G*E interactions on the prediction accuracy was also supported by the striking lack of a relationship between the mean prediction accuracy of the environments (using one environment for training) and the training population size (number of lines in the training environment) (Figure 3.2). In the absence of G*E, this relationship would be expected to be stronger. It is important to note that the matrices of prediction accuracy between environments with or without common lines were very

strongly correlated (p < 10-16). This is additional evidence that in this case the main driver of the prediction accuracy between environments was the G*E interaction. The faint triangular pattern observed in Figure 3.2 (i.e., the bottom right corner of the figure is essentially empty, whereas its other corners have observations) can be compared with Figure 6 of Lorenz et al. (2012). In that study, larger training populations were consistently more accurate whereas small training populations could sometimes achieve high accuracy presumably, just by chance. Research into methods that include G*E patterns directly into genomic selection model construction is warranted. Burgueño et al. (2012) have proposed a single-step mixed model for computing GEBV while modeling the G*E by a factor analytic model and have reported that this model increased prediction accuracies. However,myexperience with this dataset indicated that fitting such a complex model was impossible for large and unbalanced datasets, because singularity problems arose and well informed starting values had to be provided to fit the model. More research is needed to accommodate G*E in the GS models used on applied breeding data characterized by large, unbalanced datasets.

Use of genomic selection to characterize environments

This study demonstrated some of the advantages of using genome-wide marker effects to characterize environments. The use of marker effects rather than genotype performance enabled analyses to be performed even though the lines were unbalanced across environments. As genomic selection focuses on allele effects rather than on lines it also seems more appropriate to analyze the data with a direct focus on the alleles. Non-obvious outlier environments were identified without prior knowledge about each trial. This is an important feature as GS allows the use of large historic datasets where trial annotation information may have been lost, is inaccurate, or is not

readily accessible. These analyses also enabled a “forensic” approach to trial data based on which marker effects seemed particularly large in those environments, to offer hypotheses with regard to the causes making a trial exhibit outlier behavior. For the outlier environments identified based on marker effects, it is possible to examine the marker-effect distribution in that particular environment in comparison with other environments. The marker effects that are particularly large in a given environment are likely to be associated with QTL for response to the stresses experienced by lines in the environment that made it an outlier. It is important to emphasize thatmyability to detect outliers was linked to the metric used to cluster environments based on markers effect. Using a Euclidean distance caused large-effect markers to be weighted more heavily in the distance computation. However, other metrics could have also been used while searching for outliers.

For this dataset and using the Bayesian Lasso as a genomic selection model, we found that the marker effects in each environment were not dependent on the genetic composition of those environments (Table 3.1). This may be specific to this dataset because of the rather low genetic diversity of European elite, spring 2-rowed malting barley germplasm (Malysheva-Otto et al. 2006). The lack of influence of genetic composition on environment prediction accuracy facilitated the direct interpretation of the variation of marker effects among environments as a result of a differential response to the environment leading to G*E interactions.

With the growing interest in genomic selection and the concomitant emphasis on marker effects, it may be more informative to determine the ability of the trial networks to assess allelic values rather than phenotypic performance per se. Optimal trial environments for evaluating alleles may be different from those evaluating genotypes. However, the analyses performed in this study suggested that it was not the case for this dataset. In addition, the outlier environments identified by the markers

were confirmed by trial meta-data. Nevertheless, some of the outlier environments identified with the marker effects were included in the predictive set of environments, suggesting that useful information could still be extracted from some environments for GS that might be excluded for phenotypic selection. This apparent contradiction between the analyses could be reconciled when looking at the reduced kurtosis of the optimal model. It is evidence that looking for and possibly removing environments with large marker effects is beneficial to improve accuracy.

A potential challenge identified with these new methods was the impact of relatedness between environments on the prediction accuracy and on the marker-effects-based genetic distance. For this dataset, there was a significant impact of the mean kinship on the reciprocal accuracy. However, the reciprocal accuracy matrix was highly significantly correlated with the environment correlation matrix from the factor analytic model, suggesting that it captured some useful G*E signal in the complete dataset. I suggest that the use of the methods described here must be accompanied by a systematic test of the mean kinship impact on reciprocal accuracy and marker-effects clustering to guide interpretation.

Optimization of the training population for more targeted and accurate prediction

The method proposed to optimize the training population by excluding environments that are poorly predictive led to an important gain in cross-validated accuracy (from 0.54 to 0.61) while controlling overfitting, as shown by the prediction results for the new set of 2011 data. The cross-validation procedure used, by leaving one environment out and repeating the procedure, also demonstrated that it does not lead to more overfitting. This approach is difficult to validate empirically because the procedure should lead to a model optimized to predict the behavior of lines in the original TPE and it is unlikely that one year of data will be a representative sample of

the TPE. To fully validate this method, a long-term selection experiment would have to be considered.

The TPE concept is useful, but, in most cases, its specification remains elusive. A few methods have been developed to study the TPE and obtain the exact frequency of the different environment types. However, they all imply a very large research effort, not accessible to most crop breeders, or very strong and repeatable G*E patterns (Chapman et al. 2000a). Those frequencies could be derived using historic trial and weather data as described by (Löffler et al. 2005) and by using probe genotypes (Cooper and Fox 1996) to get an explicit weighting of the importance of each environment for the dataset considered (Chapman et al. 2000a; Chapman et al. 2000b; Chapman et al. 2000c).

The method we described is similar to the approach proposed for weighting environments as a function of their expected frequency in the TPE (Podlich et al. 1999), as some environments are given a weight of zero (non inclusion in the predictive set). However, instead of explicitly defining the TPE and the mixture of environments that compose it,mymethod takes a black-box approach to predicting the TPE. The assumption made is that the initial dataset is an approximate sample of the TPE. This assumption implies that the mean prediction accuracy of an environment is an indicator of the frequency of that environment in the TPE. Environments with low mean prediction accuracy are less likely to have a high frequency in the TPE. This gives us theoretical grounds to proceed with this new method.

The hypothesis that the mean prediction accuracy of an environment is an indicator of the frequency of that environment in the TPE would not hold if the dataset were comprised of several fairly distinct breeding populations that could confound line assignment to an environment with breeding populations or if the initial dataset was covering several distinct mega-environments, that is at least two distinct TPE. The

mean prediction accuracy could then be affected by the genetic relationships of the lines between environments.

The optimal model was characterized by fewer large marker effects than the model built using the complete dataset, suggesting that environment-specific QTL for responses to specific abiotic or biotic stresses that affected overall model accuracy were reduced. As an example, if one environment of a large multi-environment trial had been treated with a herbicide causing foliar damage to the crop, QTL for tolerance to that herbicide would have a strong impact on yield and might influence the model built on the whole set of environments even though the QTLs for tolerance to that herbicide have no relevance for the breeding objectives. Of course, formydata, the identification of such a stress was not as obvious as in this simple example. It is possible that for a fraction of the trials considered in such a large dataset, some of the stresses and growing conditions were not relevant to the TPE and could be detrimental to GS accuracy. In other words, it may be that the frequency of occurrence of those suppressed environments in the TPE was very low or zero. Thus, the new methodology described in this study is useful for optimization of large multi-environment trial datasets by eliminating spurious effects caused by the inclusion of low quality data. This approach can be viewed as a special case of the environment-weighting approach advocated by Podlich et al. (1999) who showed that, in the case of crossover interactions, weighting environments using their frequency of occurrence in the TPE led to a greater response to selection in the TPE. My method, by setting the weight of unpredictive environments to 0 (the unpredictive set), attempts to build a prediction model more relevant for the TPE.

Another potential explanation for the observed gain in accuracy was that this method optimized the training population for both the phenotypic records and for genetic composition. Previous empirical results showed that for some datasets there was a

wide variation in accuracy between subpopulations that could not be explained by variance heterogeneity (Heslot et al., 2012). Heslot et al. (2012) results highlighted the fact that the impact of the training population composition on accuracy was not well understood. Thus, part of the gain achieved withmymethod to optimize the training population in regards to the phenotypic data, might derive from increasing accuracy by optimization of the training population composition relative to genetic relationships of lines. However, the optimal model for this dataset retained all but one of the lines, and thus, altered training population composition did not seem to explain the realized gain in accuracy. This study demonstrates the possibilities offered by GS to analyze MET, identify outliers, group environments, and select historical data relevant for current breeding efforts.

Acknowledgments

I thank Gary Atlin for a review of an early version of this manuscript and Hugh G. Gauch for useful advice at the beginning of this project. This research was supported in part by USDA-NIFA-AFRI grants, award numbers 2009-65300-05661, 2011-68002-30029 and 2005-05130 and by Hatch project 149-449. Limagrain Europe provided financial support for N. Heslot.

References

Beeck CP, Cowling W a, Smith AB, Cullis BR (2010) Analysis of yield and oil from a series of canola breeding trials. Part I. Fitting factor analytic mixed models with pedigree information. Genome doi: 10.1139/G10-051

Burgueño J, Crossa J, Cornelius PL, Yang R-C (2008) Using factor analytic models for joining environments and genotypes without crossover genotype × environment interaction. Crop Sci doi: 10.2135/cropsci2007.11.0632

Burgueno, J., J. Crossa, J.M. Cotes, F.S. Vicente, and B. Das. 2011. Prediction assessment of linear mixed models for multienvironment trials Crop Sci. doi:10.2135/cropsci2010.07.0403

Burgueño J, de los Campos G, Weigel K, Crossa J (2012) Genomic prediction of breeding values when modeling genotype × environment interaction using pedigree and dense molecular markers. Crop Sci doi: 10.2135/cropsci2011.06.0299

Chapman SC, Cooper M, Butler D, Henzell R (2000a) Genotype by environment interactions affecting grain sorghum. I. Characteristics that confound interpretation of hybrid yield. Aust J Agric Res doi: 10.1071/AR99020

Chapman SC, Cooper M, Hammer G, Butler D (2000b) Genotype by environment interactions affecting grain sorghum. II. Frequencies of different seasonal patterns of drought stress are related to location effects on hybrid yields. Aust J Agric Res doi:10.1071/AR99021

Chapman SC, Hammer G, Butler D, Cooper M (2000c) Genotype by environment interactions affecting grain sorghum. III. Temporal sequences and spatial patterns in the target population of environments. Aust J Agric Res doi: 10.1071/AR99022

Comstock RE (1977) Quantitative genetics and the design of breeding programs. In: Pollak E, Kempthorne O, Bailey TB (eds) Proc. Int. Conf. Quant. Genet. Iowa State University Press, Ames IA, pp 705–718

Cooper M, DeLacy IH (1994) Relationships among analytical methods used to study genotypic variation and genotype-by-environment interaction in plant breeding multi-environment experiments. Theor Appl Genet doi: 10.1007/BF01240919

Cooper M, Fox PN (1996) Environmental characterization based on probe and reference genotypes. In: Cooper M, Hammer GL (eds) Plant adaptation and crop improvement. CAB Int., Wallingford, UK, pp 529–547

Cooper M, Hammer GL (1996) Plant adaptation and crop improvement. CAB Int., Wallingford, UK

Cullis BR, Smith AB, Beeck CP, Cowling WA (2010) Analysis of yield and oil from a series of canola breeding trials. Part II. Exploring variety by environment interaction using factor analysis. Genome doi: 10.1139/G10-080

Van Eeuwijk FA, Malosetti M, Yin X, et al. (2005) Statistical models for genotype by environment data: from conventional ANOVA models to eco-physiological QTL models. Aust J Agric Res doi: 10.1071/AR05153

Dekkers, JCM 2007. Prediction of response to marker-assisted and genomic selection using selection index theory. J. Anim. Breed. Genet. doi:10.1111/j.1439-0388.2007.00701.x

Elshire RJ, Glaubitz JC, Sun Q, et al. (2011) A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PLoS One 6:e19379. doi: 10.1371/journal.pone.0019379

Gilmour AR, Gogel B, Cullis BR, et al. (2009) ASREML user guide release 3.0. VSN International Ltd, Hemel Hempstead, UK

Hallauer AR, Carena MJ, Miranda Filho JB (2010) Quantitative genetics in maize breeding. Iowa State Univ. Press, Ames, IA

Hardy OJ, Vekemans X (2002) spagedi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. Mol Ecol Notes doi: 10.1046/j.1471-8286.2002.00305.x

Heffner EL, Sorrells ME, Jannink J-L (2009) Genomic selection for crop improvement. Crop Sci doi: 10.2135/cropsci2008.08.0512

Heslot N, Yang H-P, Sorrells ME, Jannink J-L (2012) Genomic selection in plant breeding: A comparison of models. Crop Sci doi: 10.2135/cropsci2011.06.0297

Holland, J.B., W.E. Nyquist, and C.T. Cervantes Martinez. 2002.Estimating and interpreting heritability for plant breeding: An update. In: J. Janick, editor, Plant breeding reviews. Vol. 22. John Wiley & Sons, Oxford, UK. p. 9–112.

Kelly AM, Cullis BR, Gilmour AR, et al. (2009) Estimation in a multiplicative mixed model involving a genetic relationship matrix. Genet Sel Evol doi: 10.1186/1297-9686-41-33

Löffler CM, Wei J, Fast T, et al. (2005) Classification of Maize Environments Using Crop Simulation and Geographic Information Systems. Crop Sci doi: 10.2135/cropsci2004.0370

Lorenz a. J, Chao S, Asoro FG, et al. (2011) Genomic selection in plant breeding : knowledge and prospects. Adv Agron doi: 10.1016/B978-0-12-385531-2.00002-5

Lorenz A. J, Smith KP, Jannink J-L (2012) Potential and optimization of genomic selection for fusarium head blight resistance in six-row barley. Crop Sci doi: 10.2135/cropsci2011.09.0503

Lorenzana, R.E., and R. Bernardo. 2009. Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theor. Appl. Genet. doi:10.1007/s00122-009-1166-3

Malysheva-Otto L V, Ganal MW, Röder MS (2006) Analysis of molecular diversity, population structure and linkage disequilibrium in a worldwide survey of cultivated barley germplasm (Hordeum vulgare L.). BMC Genet doi: 10.1186/1471-2156-7-6

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–29.

Nei M (1978) Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics 89:583–90.

Park T, Casella G (2008) The Bayesian Lasso. J Am Stat Assoc doi: 10.1198/016214508000000337

Pérez P, de los Campos G, Crossa J, Gianola D (2010) Genomic-enabled prediction based on molecular markers and pedigree using the bayesian linear regression package in R. Plant Gen doi: 10.3835/plantgenome2010.04.0005

Piepho, H.P. 1998. Empirical best linear unbiased prediction in cultivar trials using factor-analytic variance-covariance structures.Theor. Appl. Genet. doi:10.1007/s001220050885

Piepho HP, Ogutu JO, Schulz-Streeck T, et al. (2012) Efficient computation of ridge-regression best linear unbiased prediction in genomic selection in plant breeding. Crop Sci doi: 10.2135/cropsci2011.11.0592

Podlich DW, Cooper M, Basford KE (1999) Computer simulation of a selection strategy to accommodate genotype-environment interactions in a wheat recurrent selection programme. Plant Breed doi: 10.1046/j.1439-0523.1999.118001017.x

R Development Core Team (2010) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Warnes GR (2001) gplots: Various R programming tools for plotting data.

CHAPTER 4

IMPACT OF MARKER ASCERTAINMENT BIAS ON GENOMIC SELECTION ACCURACY AND ESTIMATES OF GENETIC DIVERSITY[4]

Abstract

Genome-wide molecular markers are often being used to evaluate genetic diversity in germplasm collections and for making genomic selections in breeding programs. To accurately predict phenotypes and assay genetic diversity, molecular markers should assay a representative sample of the polymorphisms in the population under study. Ascertainment bias arises when marker data is not obtained from a random sample of the polymorphisms in the population of interest. Genotyping-by-sequencing (GBS), is rapidly emerging as a low cost genotyping platform, even for the large, complex, and polyploid wheat (*Triticum aestivum* L.) genome. With GBS, marker discovery and genotyping occur simultaneously, resulting in minimal ascertainment bias. The previous platform of choice for whole-genome genotyping in many species such as wheat was DArT (Diversity Array Technology), and has formed the basis of most of my knowledge about cereals genetic diversity. This study compared GBS and DArT marker platforms for measuring genetic diversity and genomic selection (GS) accuracy in elite U.S. soft winter wheat. From a set of 365 breeding lines, 38,412 single nucleotide polymorphism GBS markers were discovered and genotyped. The GBS SNPs gave a higher GS accuracy than 1,544 DArT markers on the same lines, despite 43.9% missing data. Using a bootstrap approach, I observed significantly more clustering of markers and ascertainment bias with DArT relative to GBS. The minor allele frequency distribution of GBS markers had a deficit of rare variants compared to DArT markers. Despite the ascertainment bias of the DArT markers, GS accuracy for three traits out of fmy was not significantly different when an equal number of

[4]Heslot N, Rutkoski JE, Poland J, et al. (2013) Impact of marker ascertainment bias on genomic selection accuracy and estimates of genetic diversity. PLoS One doi: 10.1371/journal.pone.0074612

markers were used for each platform. This suggests that the gain in accuracy observed using GBS compared to DArT markers was mainly due to a large increase in the number of markers available for the analysis.

Abbreviations

BLUE, best linear unbiased estimator; BLUP, best linear unbiased predictor; DArT, diversity array technology; GBS, genotyping by sequencing; GS, genomic selection; MAF, minor allele frequency; PCA, principal component analysis; SNP, single nucleotide polymorphism;

Introduction

Genomic selection (GS) is a new marker assisted selection method based on the simultaneous use of whole-genome molecular markers to estimate breeding values for quantitative traits (Meuwissen et al. 2001). GS can accelerate the breeding cycle and increase genetic gain per unit time beyond what is possible with phenotypic selection (Heffner et al. 2010). Reviews are available on the application of GS to plant breeding (Lorenz et al. 2011).

Key to implementing GS is the availability of inexpensive whole-genome genotyping. One such recently developed platform is Genotyping-by-Sequencing (GBS) (Elshire et al. 2011). Using advances in next generation sequencing technologies, this approach uses sequencing of multiplexed, reduced-representation libraries constructed using restriction enzymes to obtain single nucleotide polymorphism (SNP) data. The multiplexed libraries are sequenced on a single run of a massively parallel sequencing platform. GBS has very low per sample costs; an ideal situation for GS in applied programs. GBS has been used with good results for GS in wheat (Poland et al. 2012b) and cassava (Ly et al. 2013). GBS has the advantage that markers are discovered using

the population to be genotyped, thus minimizing ascertainment bias. GBS typically generates a very large numbers of markers but with a high rate of missing data because genomic fragments in the library are sequenced at low depth leading to some fragments having zero coverage in some individuals.

Ascertainment bias is introduced whenever marker data is not obtained from a random sample of the polymorphisms in the population of interest. It is a sampling bias. For example, the preferential sampling of SNPs at intermediate frequencies will result in a distribution of allelic frequencies that is different compared to the expectation for a random sample. This type of biased sampling can also result from the use of a small number of lines in the SNP discovery process. This increases the frequency of the most commonly polymorphic loci and eliminates markers for loci that are less polymorphic in the screening panel. Consequently, estimates of population genetic parameters, allele frequency distribution and linkage disequilibrium can be biased (Albrechtsen et al. 2010; Nielsen and Signorovitch 2003). The effects of ascertainment bias and marker platform on genetic relationships have been studied in plants and found to have complex effects on measures of diversity and relationships between lines (Frascaroli et al. 2012; Hamblin et al. 2007; Moragues et al. 2010) that are not easily corrected.

A number of cereals are characterized by complex and large genome sizes (e.g. 16 Gb for wheat *Trititicum aestivum* L.). The predominant marker platform for whole-genome genotyping in wheat has been diversity array technology (DArT) (Akbari et al. 2006; Jaccoud 2001; Wenzl et al. 2004). DArT was developed as a hybridization-based solution, which uses a microarray platform to detect restriction site polymorphism using methylation sensitive restriction enzymes (Jaccoud 2001). DArT generates whole-genome genotypes by scoring the presence versus absence of DNA

fragments hybridized to a microarray in a reduced representation library generated from samples of genomic DNA.

DArT markers were used for most of the recent investigations concerning cereals genetic diversity and for initial studies on GS (Asoro et al. 2011; Crossa et al. 2010; Heffner et al. 2011). However, it is not known if diversity should be re-assessed using marker platforms subject to less ascertainment bias. In addition if reports suggest that GBS gives good results for GS in wheat (Poland et al. 2012b), it is not known whether that difference is due to the large increase in the numbers of markers available or to differences between the platforms. My objectives were to quantify the differences between the DArT and GBS marker platforms for population genetics metrics and GS accuracy in winter wheat, to determine if the same number of GBS markers can deliver prediction accuracies significantly different than DArT, and to determine whether any accuracy difference can be explained by either ascertainment bias, or non-random marker distribution across the genome.

Materials and methods

Data

A population of 365 soft winter wheat varieties and F5–derived advanced breeding lines originating from multiple crosses in the Cornell University Wheat Breeding Program (Ithaca, NY) was analyzed in this study. Lines were genotyped with 5,000 Diversity Array Technology (DArT) markers (Triticarte Pty. Ltd., Yaralumla, ACT, Australia), resulting in 1,544 polymorphic markers. The DArT technology for wheat assayed a reduced representation library of the genome; built on a small subset of genotypes using *PstI* and *TaqI* restriction enzymes. *PstI-PstI* fragments were cloned and the fragments polymorphic between a set of 13 Australian wheat genotypes were

printed on an array. Each clone was further validated on a large panel of genotypes for quality and polymorphism (Akbari et al. 2006; Jaccoud 2001; Wenzl et al. 2004).

All lines were genotyped using GBS as described in (Poland et al. 2012a). Briefly, after DNA digestion by two restriction enzymes, *PstI* and *MspI*, barcoded adaptors were ligated and the *PstI-MspI* fragments amplified by PCR (Polymerase chain reaction). Libraries were then pooled to 48-plex and sequenced on Illumina HiSeq2000. The sequencing reads were processed to remove potential sequencing errors and 38,412 SNPs were identified. Detailed protocols can be found in (Poland et al. 2012a) and the latest updates on the GBS approach for wheat can be found on the USDA Wheat Genetics and Germplasm Improvement website (http://www.wheatgenetics.org/research).

Phenotypic data for fmy traits were analyzed: grain yield, plant height, heading date, and preharvest sprouting (PHS) as described in (Heffner et al. 2011). Preharvest sprouting is the premature germination of seeds while still attached to the mother plant that decreases grain value and was measured as described by (Anderson et al. 1993; Munkvold et al. 2009). Phenotypic data were collected from field trials in 2008 and 2009, with three locations per year near Ithaca, NY. Each year, two locations had yield plots (1.26 m by 4 m) and one location had single 1 m rows. All traits were measured in yield trial locations, while PHS, height, and heading date were also measured in single row trials. Each location was arranged in a row-column, augmented design (Federer 1956) with six check varieties replicated 10 times each.

A two-stage analysis was used to calculate best-linear unbiased estimators (BLUEs) because it was less computationally demanding than a one-stage analysis and has been shown to generate similar results (Möhring and Piepho 2009). First, BLUEs were calculated for each trait in each location using a mixed model in ASReml-R (Gilmmy

et al. 1995). When necessary, the data was corrected for a trend along the rows and the columns of the trial and the covariance of error between neighboring plots modeled (Gilmmy et al. 1997; Malosetti et al. 2007). For PHS, an additional random effect of harvest date was included. Second, line BLUEs were calculated across years and locations. The line mean heritability was estimated to be: (yield: 0.29; heading date: 0.73; height 0.77; PHS 0.24).

Imputation of Genotypic Data

The DArT markers had 3.1% missing datapoints (cells in the marker data matrix) and the GBS data had 43.9% missing datapoints for 38,412 markers. The low level of missing data for the DArT suggested that the impact of imputation would be marginal (Rutkoski et al. 2013) for this set. Missing marker data were imputed using random forest (Breiman 2001) as described in (Rutkoski et al. 2013) separately for the DArT markers and the GBS markers. However, to be able to generate certain population genetics statistics a categorical allele call is needed. Thus, instead of random forest regression I used random forest classification to obtain a categorical allele call. Random forest is a machine-learning algorithm that uses an ensemble of decision trees, taking a majority vote of the multiple decision trees to determine a classification or a prediction value for new instances. It is a robust algorithm for classification and regression when there are thousands of input variables. In this study, a majority vote for 100 regression trees was used to impute the missing values for each marker with the RandomForest package (Liaw and Wiener 2002) in R 2.15.0 (R Development Core Team 2012) using the R package snow for parallelization. For each marker, the training set was the genotypes without missing data for that particular marker. For each classification tree, the algorithm generated a bootstrap sample as the training

population. The missing data for that marker were then predicted by each tree and the most frequently called allele was used as the imputed value.

Diversity analysis

As a first approach, principal component analysis (PCA) was used to analyze all DArT or GBS data available to look for differences in the representation of the lines. To quantify the differences between the DArT and the GBS platforms, a bootstrap procedure was used. A total of 1,544 imputed GBS markers (same as the number of DArT markers) were sampled and used to compute population genetics statistics. Based on that sample of GBS markers, lines were clustered using the R package mclust (Fraley and Raftery 2002) to identify subpopulations by hierarchical clustering using a parameterized Gaussian mixture model. The Bayesian information criterion (BIC) was used to identify the optimal number of subpopulations as well as the optimal clustering model to use. Based on that subpopulation structure, F_{st} values were computed to measure the genetic differentiation between subpopulations. F_{st} measures the fraction of the variance in allele frequencies due to population differentiation. The F_{st} estimator of (Weir and Cockerham 1984) which is insensitive to differences in subpopulation sizes was used. The overall gene diversity was computed using the R package hierfstat (Goudet 2005). To measure the information lost when the relationship matrix was calculated based on sampled GBS markers or DArT markers instead of using all GBS markers, the Kullback-Leibler divergence was used (Kullback and Leibler 1951). The Kullback-Leibler divergence is a measure of the difference between two probability distributions. It measured the information lost when the relationship based on sampled GBS markers or DArT markers are used to approximate a reference covariance matrix. The relationship matrix based on all GBS markers was used as a reference. The relationship matrix is equal to XX^t, where X is

the marker score matrix, of dimensions number of lines by number of markers. Markers are coded such that $\{aa, Aa, AA\} = \{-1, 0, 1\}$. XX^t is also referred to as the realized relationship matrix because it captures relationship between lines, including Mendelian sampling. In the context of the infinitesimal model for quantitative genetics and of genomic selection, the relationship matrix based on markers corresponds to the covariance between genotypes. For multivariate normal distribution and non-singular covariance matrix, the Kullback-Leibler divergence has a simple algebraic formulation. Calculation was carried out using the monomvn R package. The minor allele frequency (MAF) was computed for each marker. Finally, PCA analysis was carried out and the variance captured by each eigenvector calculated for each bootstrap sample. The sampling procedure was repeated 1000 times to generate a bootstrap distribution for the GBS markers.

The same statistics were computed on the entire set of DArT markers. A *p*-value was computed for the DArT value using the bootstrap GBS distribution to test the null hypothesis that, for an equal number of markers, the diversity picture is the same between GBS and DArT markers. When the *p*-value was less than 0.05, it showed the presence of significant difference between the two marker platforms for the metric considered and indicated possible ascertainment bias. Finally, PCA analysis was carried out and the variance captured by each eigenvector calculated for each bootstrap sample. The sampling procedure was repeated 1000 times to generate a bootstrap distribution for the GBS markers. The same statistics were computed on the entire set of DArT markers. A *p*-value was computed for the DArT value using the bootstrap GBS distribution to test the null hypothesis that, for an equal number of markers, the diversity picture is the same between GBS and DArT markers. If the *p*-value was significant, it showed the presence of significant difference between the two marker platforms and indicated possible ascertainment bias.

Bootstrap confidence interval

To test for significant differences between DArT and GBS platforms in terms of their MAF distributions and the percent of the variance explained by each eigenvector from PCA, a bootstrap confidence interval of the statistics of interest were calculated for the GBS marker set and then compared to that of the DArT set. Specifically, 1000 bootstrap samples of 1,544 GBS markers (same number as DArT) were drawn without replacement, and the statistics of interest were calculated. In the case of the MAF, MAF was computed for each marker and for various MAF bins the proportion of markers belonging to each bin was computed and saved for each bootstrapped sample, generating a distribution of proportions. 95% confidence intervals were then computed using these distributions and the confidence intervals were then compared to the proportion of markers in various MAF bins in the DArT marker set. In the case of percent of variance explained by each eigenvector, for each of the 1000 bootstrapped GBS samples, the percent of the variance explained by each eigenvector was calculated and saved. The distributions of these values were then compared to the percent of the variance explained with each eigenvector using the DArT marker set. Absence of overlap between the DArT values and the 95% confidence intervals of the GBS values indicated significant differences.

Redundancy analysis

A similar type of bootstrap analysis was carried out to test for a significant difference in marker redundancy. A tag SNP selection procedure (Carlson et al. 2004) was used to select one SNP for each bin of associated SNPs. Pair-wise SNP associations were measured using R^2, and SNPs within a bin that had pair-wise R^2 values greater than or equal to a specified threshold level were considered redundant. The tag SNP within a

bin was selected to minimize missing data, and there was no selection for MAF. The R^2 thresholds of 0.7, 0.8, and 0.9 were compared for the degree of redundancy. The number of tag SNPs resulting from each sample of GBS markers was computed to generate a distribution of the number of non-redundant markers for each R^2 threshold. The same tag SNP procedure was also applied to the DArT markers to estimate the number that were non-redundant at each threshold level. *p*-value s for each threshold level were computed based on the distributions of the number of non-redundant GBS markers to test the hypothesis that the number of non-redundant markers is significantly different between the DArT and GBS marker sets. As an additional test of difference in marker distribution across the genome, the variance of the Euclidean distance between markers of each GBS sample and for DArT markers was calculated. To calculate this distance between markers, the markers scores of the genotypes were used. A large variance is indicative of an uneven marker distribution across the genome and of marker clustering. A *p*-value was derived for these statistics using the bootstrap approach.

GS analysis

A bootstrap p-value was used to compare the differences in GS accuracy. For each bootstrap, 1,544 imputed GBS markers were sampled and used to compute a realized relationship matrix. A GS model was built using genomic BLUP with the R package rrBLUP (Endelman 2011). In genomic BLUP the covariance of the lines is constrained by the realized relationship matrix based on markers. A 10-fold cross validation procedure was used keeping the same partition of the folds for every bootstrap. The procedure was repeated for each of the fmy traits studied. This provided a cross-validated accuracy for each trait and each bootstrap sample. Cross-validated accuracy

was also computed using the DArT markers and a *p*-value derived using 1000 bootstrap samples.

The procedure was repeated after applying the tag SNP selection procedure to the DArT markers using an R^2 threshold of 0.8. This reduced the number of sampled markers to 787. Tag SNPs for the GBS markers were selected in the same manner, reducing the total number of sampled markers to 31,605. The 10- fold cross-validated accuracies were computed for 1000 samples of 787 GBS markers to obtain a bootstrap distribution of accuracies. The 10-fold cross-validated accuracies were also computed using the 787 non-redundant DArT markers and compared to the GBS accuracy distribution to derive a *p*-value .

Results

Diversity analysis

A population of 365 soft winter wheat varieties and F5–derived advanced breeding lines originating from multiple crosses in the Cornell University Wheat Breeding Program (Ithaca, NY) was analyzed in this study. Lines were genotyped with 5,000 Diversity Array Technology (DArT) markers resulting in 1,544 polymorphic markers. All lines were also genotyped using GBS as described in (Poland et al. 2012a). The DArT markers had 3.1% missing datapoints and the GBS data had 43.9% missing datapoints for 38,412 markers, where a data point refers to one cell in the marker data matrix. The impact of imputation on the DArT given its low level of missing data was assumed to be marginal (Rutkoski et al. 2013). Missing marker data were imputed using random forest (Breiman 2001) as described in (Rutkoski et al. 2013) separately for the DArT markers and the GBS markers. Using all the markers available for each platform revealed both similarities and differences in the Principal component analysis

(PCA) plots (Figure 4.1). Both PCA plots clearly separated the different large full-sib families present in the data. The first principle component axis explained a similar amount of variation in both analyses, and visually the relationships among lines were similar. In spite of many overall similarities, however, I detected some differences between DArT and GBS PCA. The GBS plot was rotated compared to the DArT markers plot suggesting that the second eigenvector was different between the DArT and GBS markers. There was a scale difference between the DArT markers and the GBS PCAs attributable to the large difference in the number of markers between platforms. Because there are many more markers with GBS, the distances between genotypes appeared larger.

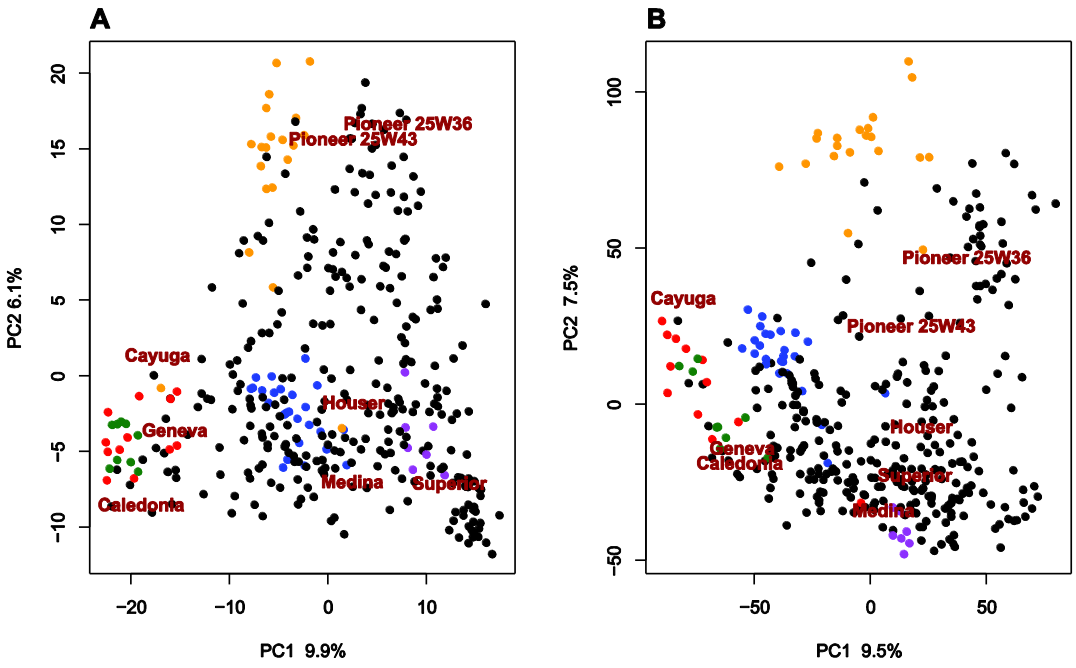


Figure 4.1. PCA plots for respectively all DArT markers (A) and all GBS markers available (B). A few large full-sibs families are color coded. (blue: Pioneer 2737W/Geneva, orange : Pioneer 2737W/Cayuga, green: Coker 8427 /AC Ron, purple: Diana/NY80095-6, red: Cayuga/Caledonia). Full-sibs are lines with the same both parents. Some of the important lines in the breeding program are indicated by their name on the plot to allow a comparison of the two PCA plots.

The small observed differences in the representation of genetic distances between lines were investigated by analyzing the R^2 between the eigenvectors of the PCA on all DArT markers with the eigenvectors of the PCA on all GBS markers. This analysis revealed some difference between platforms (Figure 4.2). If the PCA in both cases were capturing the same patterns and in the same order, the diagonal elements of the heatmap should have had a value of one and all the other cells should have had a value of 0. This was the case for the first few axes because the two first axes between both platforms were correlated (R^2 of 0.65 and 0.54 respectively). For axes three to five it appeared that both platforms captured a similar pattern but the variance was distributed differently between the axes. For the remaining axes there was very little resemblance between the patterns captured by all GBS and all the DArT markers except for the eighth principal component. This was due to ascertainment bias or to far fewer markers for the DArT markers.

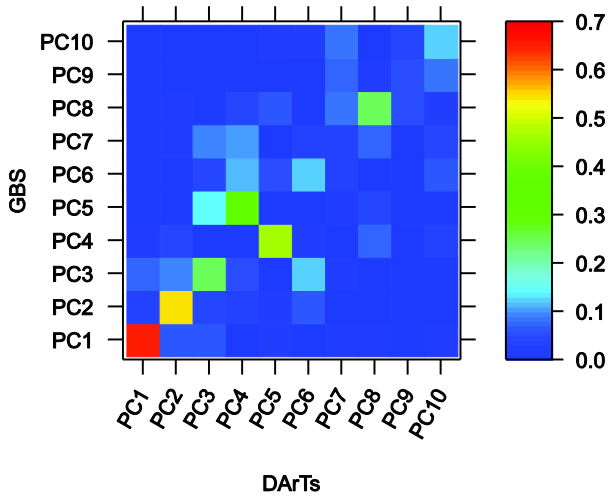


Figure 4.2. Heat map of the R^2 of the eigenvector between the two platforms. R^2 between eigenvectors of the PCA on all DArT markers and eigenvectors of the PCA on all GBS markers after random forest imputation

Those differences were further investigated and quantified by using a bootstrap approach to test significance of the differences. 1,544 GBS markers (same number as DArT) were sampled 1000 times and a number of metrics calculated. These tests were used to determine if the difference between DArT and GBS markers was significant, which would indicate that the DArT and GBS markers were drawn from different distributions.

As suggested by Figure 4.1 results, the big picture of the diversity as measured by the number of identified genotype groups was not significantly different between DArT and GBS (Table 4.1). The composition of the groups was not compared in the bootstrap approach as there was no simple statistic for comparisons with varying group numbers. Similarly, there were no significant differences in the variances explained by the first eigenvector (*p*-value 0.176) in both PCA with the bootstrap approach.

Table 4.1. Population genetics parameters computed using the DArT and p-value from the GBS bootstrap. Number of clusters of lines identified, R^2 explained by the first two PCA components, corrected for subpopulation size difference. The Kullback-Leibler divergence and the A matrix correlation test the significance of the difference between the A matrix calculated with all the GBS markers and the A matrix based on the DArT markers. Note that the bootstrap *p*-value does not compare the values obtained with all DArT markers to the value obtained with all GBS.

Parameter	**N groups geno**	**R^2 1st PC**	**R^2 2nd PC**	**F_{st} (Weir)**	**Kullback-Leibler divergence**	**A matrix correlation**
All DArT markers	8	0.099	0.061	0.24	698.32	0.7
p-value	0.205	0.176	0	0	0	0

The F_{st} (measuring sub-population differentiation) was much higher with the DArT markers than with any bootstrap sample of the GBS markers indicating a stronger

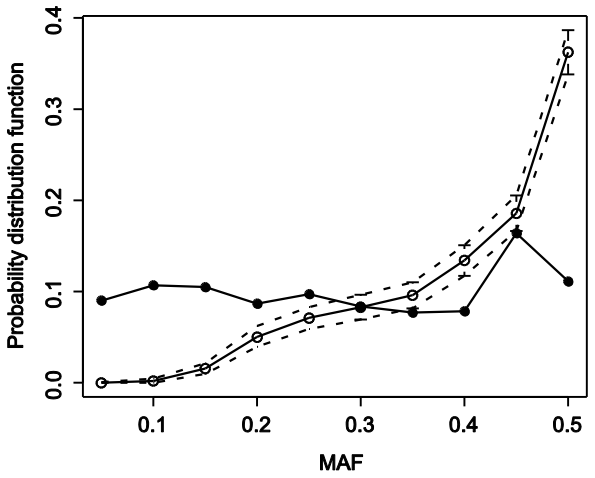
apparent population differentiation with the DArT markers. The second eigenvector of the DArT markers PCA captured much less of the total variance than any sample of the GBS markers bootstrap samples. This was an indication of an apparent more complex diversity pattern as captured by the DArT markers.

To measure the information lost when the relationship matrix was calculated using either DArT markers or an equal number of GBS markers, the Kullback-Leibler divergence was used (Kullback and Leibler 1951) with the same bootstrap approach as previously described. It measured the information lost when the relationship matrix is used to approximate a reference covariance matrix based on all the GBS markers available. The Kullback-Leibler divergence was much higher with the DArT markers than with any bootstrap sample of the GBS markers. Similarly the correlation between the relationship matrix based on DArT markers with the relationship matrix based on all the GBS markers available was much lower than with any bootstrap sample of the GBS markers. This shows that there was a significant difference in the picture of diversity captured by the two marker platforms.

The Minor allele frequency (MAF) distributions of the DArT and of the GBS markers were compared by building a 95% bootstrap confidence interval for quantiles of the GBS bootstrap distribution (Figure 4.3). If the MAF distribution for the DArT markers were not contained within the 95% confidence interval generated from the GBS bootstrap distribution, it would indicate that the MAF distribution of the DArT is significantly different from the GBS MAF distribution. The graph on Figure 4.3 shows that DArT markers are outside the GBS confidence interval for a number of MAF bins (Equal intervals of size 0.05). This indicated that the DArT markers MAF distribution significantly differs from the MAF distribution of the GBS markers. The DArT markers show a clear excess of rare variants (MAF below 0.2) compared to the GBS markers and a large deficit of frequent variants (MAF above 0.4). The MAF

distribution for all the GBS markers with and without imputation was compared, and the effect of imputation on the MAF distribution was negligible.

Figure 4.3. DArT MAF distribution and 95% confidence interval from the GBS bootstrap. The filled circle corresponds to the DArT and the empty circle corresponds to the mean of the 1000 GBS bootstrap samples.



A similar confidence interval based on a bootstrap distribution was built for the percent of variance captured by each eigenvector of the PCA and is presented in Figure 4.4. If the DArT values were outside of the 95% confidence interval it would indicate that the percent of variance captured by each eigenvector of the PCA is significantly different between GBS and DArT. The distribution of variance between eigenvectors for the DArT and the GBS markers was significantly different for every given eigenvector, except the first eigenvector (Figure 4.4). This also demonstrates that, despite an overall similar main picture (same amount of variance captured by the first component), the diversity picture was significantly different between the GBS and the DArT markers.

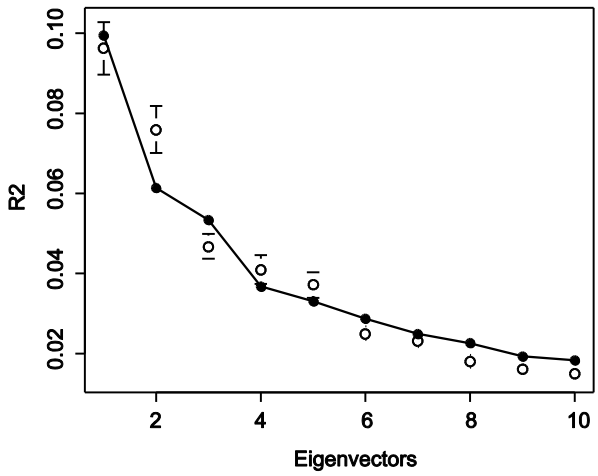


Figure 4.4. DArT PCA R^2 and 95% confidence interval from the GBS bootstrap. The filled circle corresponds to the DArT and the empty circle corresponds to the mean of the 1000 GBS bootstrap samples.

Redundancy analysis

Bootstrap p-values were calculated to test for a significant difference in marker redundancy between DArT and GBS platforms. A tag SNP selection procedure (Carlson et al. 2004) was used to select subsets of non-redundant markers. In this procedure, pair-wise marker associations were measured using R^2. The degree of redundancy in the DArT and GBS markers was assessed using R^2 cutoffs of 0.7, 0.8, and 0.9 for the tag SNP selection procedure (Table 4.2). All markers were used for this analysis.

Table 4.2. Non-redundant GBS and DArT markers and P-value function of the R^2 cutoff. Note that the bootstrap *p*-value does not compare the values obtained with all DArT markers to the value obtained with all GBS. Rather the *p*-value is for the observed DArT markers value on a bootstrap distribution of the GBS markers.

R^2 cutoff	0.9	0.8	0.7
All GBS	35,462	31,605	27,197
All DArT markers	956	787	699
P-value	0	0	0

Bootstrap p-values indicated that there were significantly more redundant DArT than GBS markers, indicating that DArT markers tended to cluster more than the GBS markers. The *p*-value corresponded to the probability of obtaining the same number or a lower number of non-redundant markers with GBS markers. Similarly, the variance of the Euclidean distance between GBS markers, calculated using their marker scores as predictors, was smaller for all GBS bootstrap samples (mean 27.61) than the DArT markers value (30.86). This indicates that GBS markers were significantly more evenly distributed across the genome than DArT markers. The same analysis was also done with non-imputed markers and gave similar results.

GS analysis

Phenotypic data for four traits were analyzed: grain yield, plant height, heading date, and preharvest sprouting (PHS) as described in (Heffner et al. 2011). Preharvest sprouting is the premature germination of seeds while still attached to the mother plant. As was done previously, a bootstrap approach was used to test for the significance of the difference in cross-validated accuracy between DArT and GBS

markers for an equal number of markers. A simple ridge regression Best linear unbiased predictor (BLUP) was used with a 10-fold cross-validation. The cross-validation partition was identical for all analyses. When using the same number of markers, redundant or not, the difference in accuracy was not significant for three out of four traits (based on bootstrap *p*-values) (Table 4.3). When using all the GBS markers available, accuracy was higher than with the DArT markers. To demonstrate that inclusion of GBS markers with high levels of missing data was appropriate, subsets of GBS markers were also selected based on a missing data threshold per marker and GS accuracies were computed. With variation between traits, accuracies reached a plateau when including markers with a high level of missing data. (Between 15% missing data for heading date and 80% for plant height). This corresponds to a minimum of 4787 GBS markers compared to 1,544 DArT markers available.

Table 4.3. Cross-validated GS accuracy for DArT and GBS and boostrap *p*-values for the DArT markers. The cross-validated accuracy is calculated using all DArT markers or all GBS or or with only the non-redundant markers, (YLD: yield, HT: height, HD: heading date, PHS: pre-harvest sprouting). P-values are presented both when all the markers were used for bootstrap and when using only the non-redundant ones for the analysis. To note that the bootstrap P-values do not compare the values obtained with all DArT markers to the value obtained with all GBS.

Trait	All DArT markers	non-redundant DArT markers	All GBS	non-redundant GBS	P-value Redundant	P-value non-redundant
YLD	0.36	0.36	0.41	0.39	0.29	0.48
HT	0.48	0.47	0.52	0.53	0.19	0.37
HD	0.30	0.31	0.47	0.43	0.22	0.56
PHS	0.47	0.47	0.57	0.56	0.00	0.06

Discussion

My analyses tested for significant differences between DArT and GBS markers when the number of markers was the same. That is, whether the DArT could have been drawn from the same distribution as the GBS markers. My results indicated that the DArT and GBS marker data yielded significantly different results for several statistics related to diversity. For a number of metrics, it was very clear that the DArT markers were not drawn from the same distribution as the GBS markers. This difference was likely due the ascertainment bias inherent in the DArT markers because DArT markers were discovered and validated on a screening panel independent from the genotyped population while with GBS the marker discovery and genotyping took place at the same time. The analyses showed that the diversity image was distorted using DArT compared to GBS markers for an equal number of markers, even though a largely similar first principle component was captured by both platforms. The difference in eigenvalues R^2 was significant between platforms indicating an apparently more complex diversity pattern as captured by the DArT markers. This would suggest that DArT markers overestimated the genetic diversity and differentiation in this population compared to the GBS markers. This was a clear indication of ascertainment bias (Nielsen and Signorovitch 2003). The significant difference in F_{st} between platforms was also an indication of ascertainment bias (Albrechtsen et al. 2010).

DArT markers had a significantly different MAF distribution from the GBS markers with an excess of rare variants compared to GBS. The different MAF distribution showed that the DArT polymorphism frequency distribution was quite different from the polymorphism frequency of all the variants in this population. This could be caused by the discovery process, done on an independent screening panel of lines. Only, polymorphisms that were in high frequency in the screening panels are genotyped, while common variants in this breeding population might have been rare or

absent in the screening panel, and thus, were not included on the DArT array. I also expect some bias with GBS. If an allele frequency is too low, it will only be read a few times, and likely be discarded by the GBS pipeline as a sequencing error.

Furthermore, I found that greater ascertainment bias in the DArT marker set led to greater redundancy of polymorphisms compared to those of the GBS marker set. This non-random sampling of polymorphisms in the genome (contributing to ascertainment bias) was most likely introduced by the restriction enzymes and screening panels used to develop the DArT array. If the restriction sites are not randomly distributed across the genome, the markers on the DArT array will also be non-randomly distributed, consistent with what I observed. DArT used *TaqI* and *PstI*, while the GBS protocol in this study used *PstI* and *MspI*. The differences between the two protocols go beyond the choice of enzymes as DArT uses arrays of cloned *PstI-PstI* fragments of size 0.4 to 1kb (Wenzl et al. 2004) while GBS directly sequences *PstI-MspI* of size between 170 and 350 bp (Elshire et al. 2011). Because of those differences in protocol it was not possible to test if the observed non-random distribution of the DArT across the genome is due to the choice of restriction enzyme itself or to other constraints of the protocol.

These findings illustrated that the reduced ascertainment bias of GBS compared to DArT markers led to differences in diversity measurements, suggesting that my knowledge of cereals diversity, which is mainly based on DArT markers, should be re-evaluated using GBS or another marker platform with reduced ascertainment bias. As no physically mapped genome sequence that is available is sufficiently anchored for wheat it was not possible to accurately assess the true distribution of polymorphisms across the genome. However, (Poland et al. 2012a) showed that the GBS markers are uniformly spaced across the genome using biparental populations. An unbiased

assessment of ascertainment bias would require knowledge of all the polymorphisms in a set of lines for a comparison to those obtained by GBS or any other genotyping method (Albrechtsen et al. 2010). Some bias might be expected of the GBS platform because the restriction enzymes usually used when creating reduced representation libraries of genomes are methylation sensitive and preferentially target gene rich regions (Elshire et al. 2011). This is potentially a problem for population genetics studies. However, at this point there is limited ability to correctly sequence and align repetitive regions such that generating markers from repetitive or gene poor regions with GBS is currently a challenge. In addition, as illustrated by Figure 4.3, identifying rare polymorphisms with GBS is currently a challenge because of confounding with sequencing errors.

Finally, despite differences due to ascertainment bias, GS accuracies between GBS and DArT markers were not significantly different for three traits out of fmy when the same numbers of markers was used. This difference was still non-significant when using sets of non-redundant markers for the DArT markers and GBS. The difference in accuracy was significant only for PHS suggesting that ascertainment bias had an impact on GS accuracy for that trait only. As DArT are not evenly spaced across the genome, they may under represent areas close to QTLs for the trait leading to a lower accuracy. When using all the GBS markers available, accuracy was higher than with the DArT markers as previously reported in (Poland et al. 2012b). This can be explained by the much larger number of markers available with GBS compared to the DArT markers. Further analysis revealed that the optimum numbers of markers varied between 4787 and 38120 GBS markers depending on the trait considered.

In terms of cost, because both platforms were designed for applications requiring high density genome coverage such as GS and association studies, the cost per genotyped entry is more relevant than cost per marker. Currently, the DArT array used here cost

approximately 50 USD per sample while the GBS protocol I used cost less than 20 USD per sample.

This study suggests that the gain in accuracy observed using the GBS compared to the DArT markers was mainly due to a large increase in the number of non-redundant markers available for the analysis. This constitutes further evidence that GBS is the marker platform of choice for further diversity analyses and GS. It also demonstrated that, given a robust imputation strategy, the high amount of missing data in GBS can be handled and imputed even without a reference map or genome sequence for application in GS as pointed out by results in Table 4.3. As SNP arrays become more widely available in wheat, it would be useful to carry out the same comparison and assess the level of ascertainment bias in SNP arrays compared to GBS. For future studies it is important to understand the quality of a genotyping platform not only based on error rate or polymorphism rate, but also based on the level of ascertainment bias and the number of non-redundant markers.

References

Akbari M, Wenzl P, Caig V, et al. (2006) Diversity arrays technology (DArT) for high-throughput profiling of the hexaploid wheat genome. Theor Appl Genet doi: 10.1007/s00122-006-0365-4

Albrechtsen A, Nielsen FC, Nielsen R (2010) Ascertainment biases in SNP chips affect measures of population divergence. Mol Biol Evol doi: 10.1093/molbev/msq148

Anderson J, Sorrells ME, Tanksley SD (1993) RFLP analysis of genomic regions associated with resistance to preharvest sprouting in wheat. Crop Sci 459:453–459.

Asoro FG, Newell MA, Beavis WD, et al. (2011) Accuracy and Training Population Design for Genomic Selection on Quantitative Traits in Elite North American Oats. Plant Gen doi: 10.3835/plantgenome2011.02.0007

Breiman L (2001) Random Forests. Mach Learn 45:5–32. doi: 10.1023/A:1010933404324

Carlson CS, Eberle MA, Rieder MJ, et al. (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am J Hum Genet doi: 10.1086/381000

Crossa J, De Los Campos G, Pérez P, Gianola D, Burgueño J, et al. (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics doi:10.1534/genetics.110.118521.

Elshire RJ, Glaubitz JC, Sun Q, et al. (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6:e19379. doi: 10.1371/journal.pone.0019379

Endelman J (2011) Ridge regression and other kernels for genomic selection with R Package rrBLUP. Plant Gen doi: 10.3835/plantgenome2011.08.0024

Federer WT (1956) Augmented (or hoonuiaku) designs. Hawaiian Plant Rec 55:191–208.

Fraley C, Raftery AE (2002) Model-Based Clustering, Discriminant Analysis, and Density Estimation. J Am Stat Assoc doi: 10.1198/016214502760047131

Frascaroli E, Schrag T A, Melchinger A. E (2012) Genetic diversity analysis of elite European maize (*Zea mays* L.) inbred lines using AFLP, SSR, and SNP markers

reveals ascertainment bias for a subset of SNPs. Theor Appl Genet. doi: 10.1007/s00122-012-1968-6

Gilmmy AR, Cullis BR, Verbyla AP (1997) Accounting for natural and extraneous variation in the analysis of field experiments. J Agric Biol Environ Stat 2:269–293.

Gilmmy AR, Thompson R, Cullis BR (1995) Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. Biometrics 51:1440–1450.

Goudet J (2005) a package for R to compute and test hierarchical F -statistics. Mol Ecol Notes doi: 10.1111/j.1471-8278 .2004.00828.x

Hamblin MT, Warburton ML, Buckler ES (2007) Empirical comparison of Simple Sequence Repeats and single nucleotide polymorphisms in assessment of maize diversity and relatedness. PLoS One 2:e1367. doi: 10.1371/journal.pone.0001367

Heffner EL, Jannink J-L, Sorrells ME (2011) Genomic selection accuracy using multifamily prediction models in a wheat breeding program. Plant Gen doi: 10.3835/plantgenome2010.12.0029

Heffner EL, Lorenz AJ, Jannink J-L, Sorrells ME (2010) Plant breeding with genomic selection: gain per unit time and cost. Crop Sci doi: 10.2135/cropsci2009.11.0662

Jaccoud D (2001) Diversity Arrays: a solid state technology for sequence information independent genotyping. Nucleic Acids Res doi: 10.1093/nar/29.4.e25

Kullback S, Leibler R (1951) On information and sufficiency. Ann Math Stat 22:79–86.

Liaw A, Wiener M (2002) Classification and regression by RandomForest. R News 2:18–22.

Lorenz AJ, Chao S, Asoro FG, et al. (2011) Genomic selection in plant breeding : knowledge and prospects. Adv Agron doi: 10.1016/B978-0-12-385531-2.00002-5

Ly D, Hamblin MT, Rabbi I, et al. (2013) Relatedness and genotype-by-environment interaction affect prediction accuracies in genomic selection: a study in cassava. Crop Sci. doi: 10.2135/cropsci2012.11.0653

Malosetti M, Ribaut J-M, Vargas M, et al. (2007) A multi-trait multi-environment QTL mixed model with an application to drought and nitrogen stress trials in maize (*Zea mays* L.). Euphytica doi: 10.1007/s10681-007-9594-0

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–29.

Möhring J, Piepho HP (2009) Comparison of weighting in two-stage analysis of plant breeding trials. Crop Sci doi: 10.2135/cropsci2009.02.0083

Moragues M, Comadran J, Waugh R, et al. (2010) Effects of ascertainment bias and marker number on estimations of barley diversity from high-throughput SNP genotype data. Theor Appl Genet doi: 10.1007/s00122-010-1273-1

Munkvold JD, Tanaka J, Benscher D, Sorrells ME (2009) Mapping quantitative trait loci for preharvest sprouting resistance in white wheat. Theor Appl Genet doi: 10.1007/s00122-009-1123-1

Nielsen R, Signorovitch J (2003) Correcting for ascertainment biases when analyzing SNP data: applications to the estimation of linkage disequilibrium. Theor Popul Biol doi: 10.1016/S0040-5809(03)00005-4

Poland J, Brown PJ, Sorrells ME, Jannink J-L (2012a) Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS One 7:e32253. doi: 10.1371/journal.pone.0032253

Poland J, Endelman J, Dawson J, et al. (2012b) Genomic selection in wheat breeding using genotyping-by-sequencing. Plant Gen doi: 10.3835/plantgenome2012.06.0006

R Development Core Team (2012) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rutkoski JE, Poland J, Jannink J-L, Sorrells ME (2013) Imputation of unordered markers and the impact on genomic selection accuracy. G3 (Bethesda) doi: 10.1534/g3.112.005363

Weir B, Cockerham C (1984) Estimating F-statistics for the analysis of population structure. Evolution (NY) 38:1358–1370.

Wenzl P, Carling J, Kudrna D, et al. (2004) Diversity Arrays Technology (DArT) for whole-genome profiling of barley. Proc Natl Acad Sci USA doi: 10.1073/pnas.0401076101

CHAPTER 5

INTEGRATING ENVIRONMENTAL COVARIATES AND CROP MODELING INTO THE GENOMIC SELECTION FRAMEWORK TO PREDICT GENOTYPE BY ENVIRONMENT INTERACTIONS[5]

Abstract

Genotype by environment interaction (G*E) is one of the key issues when analyzing phenotypes. The use of environment data to model G*E has long been a subject of interest but is limited by the same problems as those addressed by genomic selection methods: a large number of correlated predictors each explaining a small amount of the total variance. In addition, non-linear responses of genotypes to stresses are expected to further complicate the analysis. Using a crop model to derive stress covariates from daily weather data for predicted crop development stages, I propose an extension of the factorial regression model to genomic selection. This model is further extended to the marker level, enabling the modeling of quantitative trait loci (QTL) by environment interaction (Q*E), on a genome wide scale. A newly developed ensemble method, soft rule fit, was used to improve this model and capture non-linear responses of QTL to stresses. The method is tested using a large winter wheat dataset, representative of the type of data available in a large-scale commercial breeding program. Accuracy in predicting genotype performance in unobserved environments for which weather data was available increased by 11.1% on average and the variability in prediction accuracy decreased by 10.8%. By leveraging agronomic knowledge and the large historical datasets generated by breeding programs, this new model provides insight into the genetic architecture of genotype by environment interactions and could predict genotype performance based on past and future weather scenarios.

[5] Heslot N, Akdemir D, Sorrells ME, Jannink J-L (2013) Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions doi: 10.1007/s00122-013-2231-5 with kind permission from Springer Science+Business Media B.V

Abbreviations

BLUP, best linear unbiased predictor; GBLUP, genomic estimated best linear unbiased predictor; GEBV, genomic estimated breeding value; G*E, genotype by environment interactions; GS, genomic selection; MET, multi-environment trials; QTL, quantitative trait locus; Q*E, QTL by environment interaction; SGL, sparse group lasso; SNP, single nucleotide polymorphism; TPE, target population of environments

Introduction

Genotype by environment interactions (G*E) are one of the most important issues in plant breeding. It is a frequent observation in multi-environment trials (MET) that genotype performance varies across environments leading to variance differences and rank changes among genotypes (Cooper and DeLacy 1994). These forms of G*E are called non-cross-over (variance differences) and cross-over G*E (rank changes). Rank changes complicate selection for broad adaptation as there might not be one best performing genotype everywhere. In a context of climate change and reduced usage of fertilizers and pesticides, crop environments will likely be more variable, increasing the importance of G*E.

Genomic selection (GS) was first proposed by Meuwissen et al. (2001) to better estimate breeding values based on the simultaneous use of whole genome markers. A number of empirical and theoretical studies suggest that it could increase genetic gain per unit of time beyond what is possible with phenotypic selection (Heffner et al. 2010; Lorenz et al. 2011).

The GS concept provides an important breakthrough towards a better genotype to phenotype mapping. It solves problems encountered in the application of quantitative trait loci (QTL) study results to breeding, such as overestimation of the identified QTL

effects. GS also potentially enables the use of historical breeding data for current breeding efforts. However, an important part of the mapping, the differential response of genotypes to the environment, which causes G*E, (Van Eeuwijk et al. 2005) has yet to be included directly into GS approaches. Genomic predictions have so far focused on the computation of breeding values that are single point estimates of genotype performance presumed to be useful across all environments.

In the context of classical plant breeding the G*E issue has been tackled in several ways (DeLacy et al. 1996), the most common is to ignore G*E in the analysis by considering it to be noise. Another approach is to identify repeatable G*E patterns in the data by dividing the environment targeted by breeding into mega-environments that minimize G*E within mega-environments. This allows genotype targeting and increases the trait heritability within the mega-environments, provided sufficient breeding resources are allocated to each mega-environment (Windhausen et al. 2012). Numerous approaches have been developed to group environments such as AMMI (Gauch 2006), and clustering (Cooper and DeLacy 1994). Repeatable G*E patterns can also be identified using external data (Löffler et al., 2005; Chenu et al. 2013). For example, if drought stress is known to be the main driver of G*E, environments can be clustered based on drought patterns (Chapman et al. 2000b).

The most powerful integration of G*E within quantitative genetics theory is to consider G*E as a lack of genetic correlation between environments (Falconer and Mackay 1996). Genetic correlations between environments are obtained by considering performances in different environments as different correlated traits.

When G*E is considered as a lack of correlation, it can be taken into account using multiplicative mixed models such as the factor analytic structure to model the covariance between environments responsible for G*E (Piepho 1998; Burgueño et al. 2008; Kelly et al. 2009; Cullis et al. 2010). Those models can be used for GS

prediction, by using a relationship matrix based on markers in the mixed model (Burgueño et al. 2012). In practice, those approaches have numerical limitations due to the highly unbalanced nature of most multi-environment plant breeding datasets. In addition, because they are based on observed covariance among environments, they are explanatory a posteriori rather than predictive. They don't allow prediction for a new climatic scenario, a given level of a weather-related stress or a new environment directly.

A way to gain predictive capability of G*E is to investigate the genetic basis of G*E by identifying the environment parameters responsible for G*E and determining genotype sensitivity. The class of models implementing this approach is termed factorial regression (Denis 1988; Piepho et al. 1998). Genotype performances are the sum of the main genotype effect and of the genotype sensitivity to the stress covariate. Factorial regression has been extended to the differential response of QTL in the biparental mapping case for a few detected QTL (Crossa et al. 1999; Malosetti et al. 2004; Boer et al. 2007).

Including stress covariates in the analysis presents some of the same issues encountered using genomic selection (GS) methods for estimating breeding value. A very high number of covariates can potentially be obtained, each explaining a small amount of the total variance while being highly correlated with each other (Brancourt-Hulmel et al. 2000). Leo Tolstoy's Anna Karenina provides a metaphor for the difficulty of modeling G*E: "Happy families are all alike; every unhappy family is unhappy in its own way." In the context of crops, when genotypes perform poorly in a given environment it can be due to many different stresses and deriving general results is a daunting task. On an operational level, the Anna Karenina effect occurs when most of the G*E can't be explained by a few major stresses or a simple geographic partition of the data.

To be successful for prediction, an approach focusing on the genetic basis of G*E would have to be genome-wide and include numerous stress covariates at the same time. Because response and development curves are often exponential or 'S' shaped (Van Eeuwijk et al. 2005) the framework should accommodate non linear responses of QTL to the environment variables. Such a framework should enable the modeling of G*E at the allele level focusing on QTL by environment interaction (Q*E).

Considering those challenges, some groups have focused on crop modeling to better understand G*E and incorporate external information about the crop (Chenu et al. 2008; Messina et al. 2009). Crop models are sets of equations developed by extensively studying the behavior of a few genotypes under a range of growing conditions. Their main purpose is to predict the development of a crop and the genesis of the different yield components. Figure 5.1 presents the schematic structure of a crop model. Crop models were initially developed to assist in crop management decisions, strategic planning, yield forecasting, and definition of research needs.

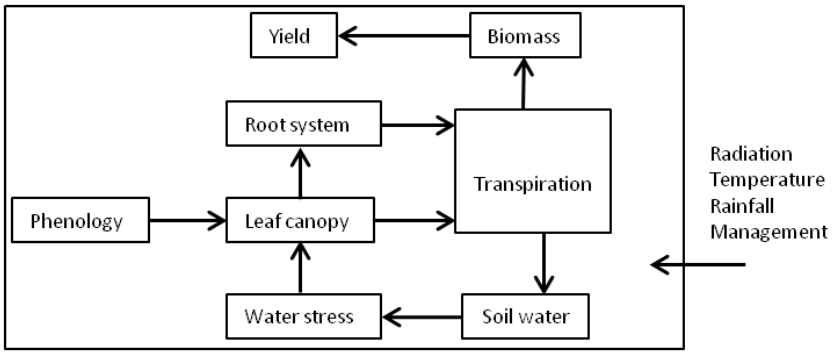


Figure 5.1. Typical structure of a crop model

By design, a crop model integrates environment inputs such as weather and soil data in a non-linear way. Crop models, however, have been criticized by crop geneticists for not taking enough genetic variation into account (Hammer et al. 2002). To use crop

models for trait prediction across genotypes, research focuses on variation in the parameters of the crop model (Quilot et al. 2004; Jullien et al. 2011). Those parameters are expected to be more heritable than the trait predicted by the crop model, less prone to G*E and to have a less complex genetic architecture. It is a very appealing strategy but it also requires the measurements of specific phenotypes to recover the model parameters. This is not trivial as large numbers of environments have to be sampled to study G*E. Another possibility is to use crop modeling as a tool to perform a physiological integration of environmental data in order to derive stress covariates (Landau et al. 1998; Landau et al. 2000; Boer et al. 2007). These covariates are then used as independent variables in quantitative and statistical genetic models for effect estimation and prediction. This has the advantages of using a genetic model to predict the main genotype effect, whose optimality properties are well known. Using a crop model to parameterize the environment data reduces data dimensions from daily weather variables to a few covariates per crop growth stage. The daily weather data is composed of numerous correlated variables most of which have little or no impact on the crop. Moreover, there is a wealth of agronomic and physiological knowledge about the sensitivity of specific growth stages to specific abiotic stresses (Meynard and Sebillotte 1994) (Figure 5.2). In addition, the use of a crop model to predict development stages may capture part of the non-linear response of genotypes to the environment by modeling non-linear development processes such as vernalization. It also eliminates the need for specific experiments to measure crop model parameters, and thus enables the use of large commercial breeding datasets that often contain no more than measurements on the final trait of interest, yield. This makes the assumption that the stress response genetic architecture is the same among genotypes at a given developmental stage. Furthermore, interpretability of the model is improved because it uses stress covariates defined by growth stage.

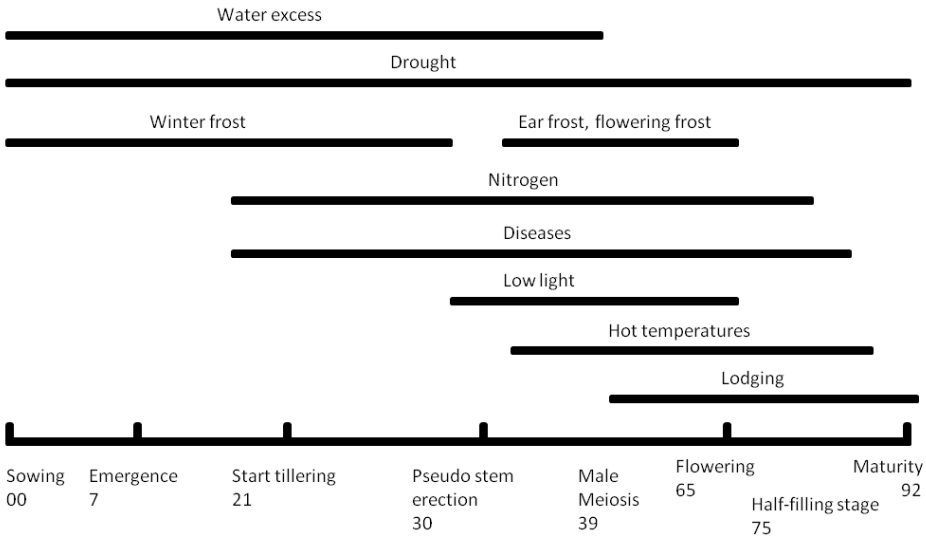


Figure 5.2. Major stresses by development stage for winter wheat from Meynard and Sebillotte (1994). The numbers below the development stage names correspond to the Zadoks scale (Zadoks et al. 1974)

The broad objective of my research is to propose new solutions to integrate environmental data and crop modeling into the genomic selection framework to predict G*E. To this end, I explicitly model whole-genome markers and their differential response to the environment in the GS context to better understand the genetic architecture of G*E. In this study, I extended factorial regression to the GS context and developed a new machine learning approach to capture the response of QTL to stresses non parametrically. This approach was used along with a crop model to enable the use of daily weather data in prediction models. Those G*E predictions could be used to make breeding decisions for specific adaptation. However, as presented in the discussion I believe that they are more useful as a tool to understand the interaction and the structure of the target population of environments (TPE). The TPE is the mixture of environments expected for the intended target region (Comstock 1977). This information could then be used to optimize phenotyping.

Materials and methods

Phenotypic and environment data

Limagrain Europe (Chappes, France) provided a large commercial winter wheat breeding dataset to assess the predictive power of the new models. It consisted of 2,437 genotypes tested for grain yield in 12 locations across France from 2006 to 2011 for a total of 44 environments (year-location combinations). The total number of yield plots was 23,265 generating 9,024 within-environment adjusted genotype means. Those environment means accounted for experimental design which varied from complete block to alpha-lattice and were corrected for spatial variation. The numbers of genotypes observed in each environment ranged from 52 to 974. The data was generated by a single breeding program and corresponded to all trials for the three first years of yield testing of genotypes. As the genotypes were advanced in the breeding program the level of replication and the number of locations increased while some genotypes were discarded based on their past performance. The dataset was unbalanced with non-random missingness of the genotypes due to the historical nature of the dataset and because France can be divided in two target environments for winter wheat, the South favoring early maturing genotypes to escape drought and high temperature and the North favoring late maturing genotypes that yield more. Genotypes were dropped from testing based on their previous performance and new ones added over time. My data is similar to the winter wheat example in (Piepho & Möhring 2006): Because an important focus of the breeding is quality, which tends to be negatively correlated with yield, the genetic variance for yield should not decrease dramatically over time. They did not find strong evidence of downward bias for variances of genotype main effects and genotype by year interactions. On average each genotype was observed in 3.7 environments with 760 genotypes observed in only

one environment. The genotypes and the environments corresponded to only one breeding program. This means that any G*E present in this dataset is considered to be small enough that it is manageable by having only one breeding program. Trials were conducted using standard agronomic practices including appropriate use of fertilizers and fungicides.

The lines were genotyped with 1,287 SNP and haplotype-based markers covering the whole genome. Some markers provided perfect linkage with major adaptation loci (dwarfing genes: *Rht-B1*, *Rht-D1*, vernalization genes: *Vrn-A1*, *Vrn-A2*, *Vrn-D5*, *Vrn-B3*, photoperiod sensitivity genes: *Ppd-D1*, *Ppd-B1*, *Ppd-A1*).

In addition, latitude and longitude of each trial location were available along with the sowing date. Daily weather data were obtained from the AGRI4CAST action of the Joint Research Center of the European Commission. (http://mars.jrc.ec.europa.eu/mars/About-us/AGRI4CAST/Introduction). Those data are used to generate crop yield forecasts for European policy makers. The database contains daily meteorological data from 1975 to the last calendar year completed, covering the European Union and neighboring countries. The meteorological parameters are interpolated to a 25x25 km grid from a network of meteorological stations. Details about the interpolation procedure and calculations can be found in Van der Goot and Orlandi (2003). The variables available were the mean, minimum and maximum daily temperature, daily precipitation, daily global radiation, and the ETP (Penman potential evapotranspiration from a crop canopy (mm/day)). The quality of the data was verified using independent temperature and rainfall records obtained from two trial sites over several years. The interpolated data were well correlated with the observed data for temperature, (correlation above 0.9), less so for rainfall with a correlation of 0.6 going up to 0.8 when considering a weekly scale.

Derivation of stress covariates from the weather data

An intuitive approach to include weather data in an interpretable way, to reduce the number of variables and accommodate some non-linearity of response was to define stress covariates by development stage (Landau et al. 1998; Landau et al. 2000; Boer et al. 2007). Brancourt-Hulmel et al. (1999, 2000) and (Lecomte 2005) developed a set of stress indices for winter wheat. In their studies they determined stress covariates by analysis of yield components such as thousand kernel weight and number of kernels per surface area, to identify yield-limiting factors per stage. In addition, they compiled winter wheat sensitivity to stresses at specific development stages from previously published work. Figure 5.2 shows which stresses are expected to occur for each development stage.

Some of the original stress covariates were excluded because of the lack of necessary information to compute them. They included stress covariates accounting for winter frost damage, disease pressure and nitrogen availability.

Table 5.1. Stress covariates used and references, modified from Lecomte (2005) page 87. (-w: winter period - sowing to 1cm-ear stage, -em: 1cm ear stage to meiosis, -mf: meiosis to flowering, -f30: flowering–30 days to flowering, -fh: flowering to half filling stage, -hm: half filling stage to maturity; P: rainfall in mm, ETP: potential evapotranspiration in mm, dd: degrees days)

Abbreviation	Description
stmpw, stmpem, stmpmf	Sum of the daily average temperatures (°C) above 0 by development periods
sradw, sradem, sradmf, sradfh, sradhm	Sum of the daily radiation (J/cm²) by development periods (Gallagher and Biscoe 1978; Monteith 1972)
rdtmpw, rdtmpem, rdtmpmf, rdtmpf30	Ratio srad / stmp by development periods (Fischer 1985)
watxw	Sum of the daily differences P-ETP (mm) >0 from sowing to pseudo stem erection
spetpw, spetpem, spetpmf, spetpfh, spetphm	Sum of the daily differences P-ETP (mm) <0 by development stages
spetpe1	Sum of the daily P-ETP (mm) from pseudo stem erection minus 150dd to pseudo stem erection plus 350dd
nsddr	Number of successive dry days (P<=ETP in mm) from pseudo stem erection -150dd to pseudo stem erection +350dd
ntddr	Number of total dry days (P<=ETP in mm) from pseudo stem erection stage minus 150dd to pseudo stem erection plus 350dd
ndefr	Number of days of ear frost (minimal temperature <=-4°C) from pseudo stem erection to flowering (Gate 1995)
sti4	Sum of the daily minimal temperatures <-4°C from pseudo stem erection stage to flowering
ndt0f	Number of days when the daily minimal temperature is <=0°C from heading to heading plus 300dd
st0f	Sum of the daily minimal temperatures <0° from heading to heading+300dd
sradmg, sradmm	Sum of the daily radiation (J/cm²) from meiosis-100dd to heading, or from meiosis-5d to meiosis+5d (Demotes-Mainard et al. 1996)
ndi10m	Number of days when the radiation is <=1045 J/cm² from meiosis minus 5 days to meiosis plus 5 days
sri10m	Sum of the daily radiation <1045 J/cm² from meiosis minus 5 days to meiosis plus 5 days
nd25m	Number of days when the maximal temperature is >=25°C meiosis minus 5 days to meiosis plus 5 days
st25m	Sum of the daily maximal temperatures >25°C from meiosis minus 5 days to meiosis plus 5 days
nd25ef	Number of days when the maximal temperature is >=25°C from heading to flowering (Tashiro and Wardlaw 1990)
st25ef	Sum of the daily maximal temperatures >25°C from heading to flowering
nd25fh	Number of days when the maximal temperature is >=25°C from flowering to half-filling stage (Hunt 1991; Sofield et al. 1977; Stone and Nicolas 1998)
st25fh	Sum of the daily maximal temperatures >25°C from flowering to half-filling stage
nd25hm	Number of days when the maximal temperature is >=25°C from half-filling stage to maturity
st25hm	Sum of the temperatures maximal daily >25°C from half-filling stage to maturity

Use of a crop model to predict development stages

To derive the stress covariates from the daily weather data, the development stage timing has to be known. This information is often difficult to obtain or not available. To alleviate that need, a crop model, SiriusQuality, was used (Martre et al. 2006). This model is process-based and used a modified version of the phenology model proposed by (Jamieson et al. 1998).

The daily weather data was retrieved from the database using the longitude and latitude of the trial locations. Those data were used as input parameters, to obtain development stages for the crop model SiriusQuality, with default parameters for non-limiting water and nitrogen. The stages predicted by the model used the Zadoks scale code (Zadoks et al. 1974) and were stages 30 (Pseudo stem erection), 39 (Flag leaf ligule just visible, male meiosis), 65 (Anthesis), 75 (Half-filling stage), 92 (Maturity). The calendar date of stage 55 (Heading date) was derived from the daily sum of temperatures taking stage 39 (male meiosis) as a reference (Gate 1995): First *stepi* , the sum of daily mean temperature in base 0°C from planting to heading was calculated using the daily sum of temperature in base 0°C from planting to meiosis *stmei* obtained from the crop model as $stepi = (stmei - 74)/0.864$. Then *stepi* was converted into a calendar date.

Ideally, the growth stages of each genotype in each environment should be obtained or computed. As this was not feasible and the data covered a wide range of maturities, three sets of development stages were obtained for three elite genotypes, Soissons, Thésee, and Renan, which are early, mid and late maturing, respectively (He et al. 2012). Those three genotypes were all commercially successful at some point in the last 30 years.

Once the development stages are known, the daily weather data can be used to compute the stress covariates described in Table 5.1. For example, for the stress

covariate stmpmf, the average daily temperature when it is above 0°C is summed between the predicted meiosis and flowering date and this for the three sets of development stages. The covariates capturing frost stress in the spring (ndefr, sti4, ndt0f, st0f) were removed because they had no variance. The three sets of stress covariates plus latitude and longitude of environments were used together (for a total of 101 covariates) and were standardized to zero mean and unit variance before further use. The whole set of covariates (101) was used for all the genotypes regardless of their maturity. This means that for each stress covariate and each observation there are three values corresponding to the three genotypes used to parameterize the crop model.

Mixed model formulation of G*E

The simplest model to analyze multi-environment trials, in the balanced case, with one observation per genotype and environment combination, for m genotypes and t environments is (Piepho et al. 2008):

$y = 1_{mt}\mu + (I_t \otimes 1_m)\beta + (1_t \otimes I_m)u + \varepsilon_1$ (Model 1)

Where $\otimes$ indicates the Kronecker product, y, observed phenotypes, β, effects due to the design such as environment in the case above (although block effects could also be represented) and generally treated as fixed, and u are random line effects. In the classic one-step GS approach the covariance of u is $A\sigma_u^2$ with A the realized relationship matrix computed using molecular markers as $A = VV^t$ with V the marker score matrix. V has dimensions $m \times n$, n numbers of markers. Markers are coded $\{aa, Aa, AA\} = \{-1, 0, 1\}$. σ_u^2 is the genetic variance to be estimated, for example with restricted maximum likelihood. This approach is often referred to as GBLUP for Genomic BLUP. To prevent singularity issues in A the relationship matrix was bended by adding a small scalar to the diagonal elements (Piepho et al. 2012). ε_1 is

assumed to be normally i.i.d. Homoscedasticity is probably an incorrect assumption here. However, when the number of replicates per adjusted mean was used to weight the observations, no gain in accuracy was observed and accuracy even slightly decreased. This is attributable to the variety of experimental design used for the trials forming this dataset. (RCB, alpha-lattice, unreplicated designs). It suggests both that there is no easy way to adjust for heteroscedasticity here and that I pay no penalty for the simplifying homoscedasticity assumption. In the following models I assume that this assumption holds. Here the G*E variance is absorbed by the residual ε_1. Habier et al. (2007) has shown the equivalence of GBLUP and ridge regression. That model can then be equivalently written:

$y = 1_{mt}\mu + (I_t \otimes 1_m)\beta + (1_t \otimes V)\alpha + \varepsilon_2$ (Model 2)

Where α is a marker effect vector: u from Model 1 is equal to $V\alpha$ here.

G*E can be accounted for explicitly by the following model:

$y = 1_{mt}\mu + (I_t \otimes 1_m)\beta + (1_t \otimes I_m)u + (I_t \otimes I_m)ge + \varepsilon_3$ (Model 3)

With ge a vector of G*E deviations with covariance proportional to $I_t \otimes A$. However, this model does not provide for a detailed analysis of G*E. To integrate G*E in the mixed model analysis, it is most straightforward to consider G*E as a lack of genetic correlation between environments (Falconer & Mackay 1996). Then, performances in different environments are different traits. This model can be written as

$y = 1_{mt}\mu + (I_t \otimes 1_m)\beta + (I_t \otimes I_m)\eta + \varepsilon_4$ (Model 4)

Where η is the vector of environment-specific effects for each genotype with covariance $G \otimes A$. G is the covariance matrix of genotype effects in environments with dimensions $t \times t$. G can be estimated using a factor analytic model for example (Piepho 1998; Burgueño et al. 2008; Kelly et al. 2009; Cullis et al. 2010). However Model 3 and Model 4 do not allow prediction of G*E in unobserved environments.

Introducing in the model differential genotype sensitivity to specific stress covariates can capture part of the G*E interaction and allow prediction of performance in unobserved environments. This class of model is termed factorial regression model (Denis 1988; Van Eeuwijk et al. 1996).

For the balanced case, the model can be written in matrix notation:

$$y = 1_{mt}\mu + (I_t \otimes 1_m)\beta + (1_t \otimes I_m)u + (S \otimes I_m)\gamma + \varepsilon_5 \quad \text{(Model 5)}$$

Where S is a matrix of dimensions $t \times q$ which contains the centered and scaled observed scores of the q stress covariates in each of the t environment, and γ is the vector, mq long, of stress-specific sensitivities for each genotype with covariance $H \otimes A$, where H is a $q \times q$ covariance matrix of the stress covariates. The definition of an adequate covariance structure for γ is a problem noted in the literature (Smith et al. 2005). Here, the GBLUP framework was extended to factorial regression. This gave additive genotype sensitivity to any given covariate, which is of interest for breeding as well as providing a way to cope with unbalanced data.

Model 5 can be seen as a random regression model, with u random intercept for each genotype and γ genotype specific random slope. Then, the assumption that the average of the regression slopes across genotypes is zero is not realistic. The assumption could be relaxed by including a fixed effect for each stress covariate. However, here the stress covariates are confounded with the environments such that the environment effect β captures the mean regression of genotypes on the stress covariates.

For the Model 5 to be scale invariant to linear transformations of the stress covariates, it would be necessary to include a covariance term between u and γ, which would further increase its complexity. The lack of scale invariance means that the model is limited when making inferences about the variance components or testing for fixed effects. The limitation noted for example, in Smith et al. (2005) remained, as often any given stress covariate explains only a small proportion of the G*E interaction and a

large number of variance components has to be estimated. Considering an equivalent model at the marker level instead of the genotype was investigated as a way to simplify Model 5.

Factorial regression at the QTL level

The factorial regression framework was extended at the QTL level, by modeling each QTL effect as a combination of a main effect and a function of the stress covariates (Crossa et al., 1999; Malosetti et al., 2004), with application in Boer et al. (2007). The design matrix for the linear sensitivity of the markers to each of the stress covariates is in the balanced case: $S \otimes V$, with V the design matrix for all n markers. Combining Model 2 and Model 5, I obtain:

$$y = 1_{mt}\mu + (I_t \otimes 1_m)\beta + (1_t \otimes V)\alpha + (S \otimes V)\varphi + \varepsilon_6 \quad \text{(Model 6)}$$

Where α is a marker effect vector, and φ is a vector of linear sensitivities of the markers to the stress covariates. Note that φ contains $q \times n$ parameters, so that this model has high dimension. Model 6 can be interpreted as a large penalized regression. However, different levels of shrinkage are desirable for each group of variable (α, β, φ).It is expected that the φ (marker sensitivities to the environment) would have to be shrunk more than the α (main marker effects). One way to solve this problem in a single step analysis would be to use the sparse group lasso (SGL) (Friedman and Hastie 2010) and defining groups of variables for α, β and φ to provide differential shrinkage of the different groups of variables and enforce model sparseness.

Two-step approach

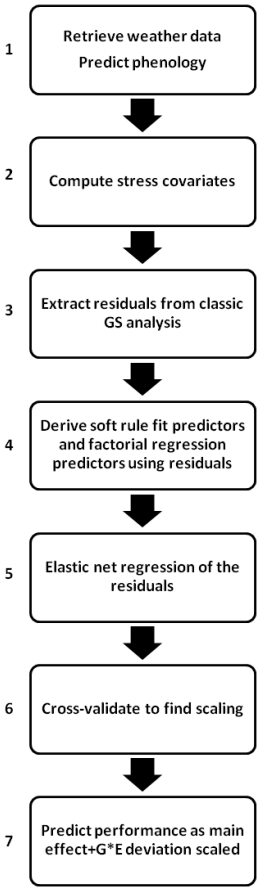


Figure 5.3. Modeling flow diagram

Algorithms for the SGL are not as computationally efficient as lasso methods, and this became a major hurdle when dealing with thousands of predictors and thousands of observations. Initial testing with the R package scoop (Chiquet et al. 2012) showed that the SGL was too slow to be a practical method with a dataset as large as mine. I therefore adopted a two-step approach where the main marker and environment effects were first computed using the Bayesian Lasso (Park and Casella 2008) implemented in the R package BLR (Pérez et al. 2010). Residuals from the main effect model (Model

2) and the corresponding β and α were extracted. The residuals of Model 2 ε_2 are deviations from an additive model. They contain G*E deviations in addition to random error. This residual extraction corresponded to step 3 of Figure 5.3. At this point there is only one group of effects, φ , to be estimated and this can be done using a simple penalized regression method. To gain full equivalence with a single-step analysis, the residuals ε_2 should be regressed not on the predictors for φ but on predictors corrected for the main marker effects. However, Model 6 is not a simple multiple regression model, because the factorial regression model involves only a regression of the residual from additivity on the environmental covariates. As a consequence it does not seem strictly necessary to correct the predictors for φ. For step 4 of Figure 5.3, the predictors for the $(S \otimes V)\varphi$ term in (Model 6) were regularized on the residuals of the main effect model with an elastic net (Zou and Hastie 2005). It relies on a combination of both the L_1 (lasso) and L_2 (ridge) norm penalties. For the regression problem $\varepsilon_2 = \mu + (S \otimes V)\varphi + \varepsilon$, The estimator of φ is

$$(1+\lambda_2)\arg\min_{\varphi}\left(\left[\varepsilon_2 - (S \otimes V)\varphi\right]'\left[\varepsilon_2 - (S \otimes V)\varphi\right] + \lambda_2 \|\varphi\|^2_2 + \lambda_1 \|\varphi\|_1\right),$$

with λ_1 and λ_2 shrinkage parameters. This was implemented using the R package glmnet (Friedman et al. 2010). The shrinkage coefficients were estimated from the data using cross-validation. Fitting these residuals corresponded to step 5 in Figure 5.3. Using elastic net had the advantage of avoiding strong assumptions about the optimal model sparseness between a lasso and a ridge regression. For the elastic net properties to hold, all predictors $S \otimes V$ were centered and scaled prior to the analysis. Because the prediction for the markers main effect α and the predicted G*E deviation $V\varphi$ were both generated by penalized regression methods, they had to be rescaled before combining them to predict the phenotype. Model 6 can then be rewritten as:

$$y = 1_{mt}\mu + (I_t \otimes 1_m)\beta + (1_t \otimes V)\alpha + \theta(S \otimes V)\varphi + \varepsilon_7 \quad \text{(Model 7)}$$

Where θ is a scaling parameter which was determined by cross-validation within the training set to maximize phenotypic prediction accuracy within environments. This optimal θ was further used with the full model fitted on the whole training set to predict genotype performance in unobserved environments. Thus, the phenotypic performance in environment i was predicted as $V\alpha + \theta\left(b_i^t S\right) \otimes V\varphi$. b_i is a column vector with the ith element equal to one and zero elsewhere. This two-step approach is very similar to the back-fitting algorithm for generalized additive models proposed by (Breiman and Friedman 1985). It was also suggested by (Gianola et al. 2006) for combining a classic additive model with a non-parametric component in a mixed model to predict for example milk production in cows. It involves several iterations of the process of fitting the model terms sequentially on the residuals of the previous terms. This is in essence the procedure I use here.

Selection of a marker subset to use for factorial regression

Fitting linear marker sensitivity to each covariate would require as many predictors per marker as there are covariates. To reduce the dimensionality of the problem, a subset of markers was selected as follows. In each environment, marker effects were computed separately as in (Heslot et al. 2013). Using the Bayesian Lasso (Park and Casella 2008) implemented in the R package BLR (Pérez et al. 2010), the model was run for 60,000 iterations and the first 20,000 were discarded as burn-in and the chains were not thinned. Model convergence was visually assessed based on the trace of parameter samples over iterations.

The variance of marker effects across environments was computed based on the table of marker effects in each environment. Markers were ranked accordingly and sets of different size s of the most variable markers across environments were used to build predictors for the factorial regression at the marker level. This is an important

difference from previous approaches taken for example in Boer et al. (2007) as they constructed factorial regression predictors at the marker level only for the QTL detected in at least one environment. The optimal number of markers s to be included in the model was determined by cross-validation. Model 7 then became:

$y = 1_{mt}\mu + (I_t \otimes 1_m)\beta + (1_t \otimes V)\alpha + \theta(S \otimes V_s)\varphi + \varepsilon_8$ (Model 8)

V_s is a subset of V the marker design matrix, containing a subset of s markers selected as described above.

An ensemble method to model complex responses of QTLs to stresses.

An important limitation of the factorial regression method is the difficulty to properly model non-linear responses of QTL to stress covariates (Van Eeuwijk et al. 2005). To my knowledge this modeling has only been attempted in biparental QTL mapping and with one covariate, (Ma et al. 2002).

From physiology knowledge, Model 7 is expected to be inadequate because it is linear but the relationship between the response and the predictors is expected to be non-linear. The response could be approximated using polynomials or splines but this would require a very large number of predictors and is impractical. In a more general case Model 8 can be rewritten as:

$y = 1_{mt}\mu + (I_t \otimes 1_m)\beta + (1_t \otimes V)\alpha + \theta f(S,V)\xi + \varepsilon_9$ (Model 9)

Here the G*E response is determined by a function $f(.)$ which depends on the genotype and the stress covariates. ξ corresponds to effects of the predictors from $f(S,V)$. Model 9 reduces to Model 8 by setting $f(S,V)\xi = (S \otimes V_s)\varphi$. This function is expected to be complex but could be approximated using machine learning techniques suited for non-linear problems. It is also expected that linear response of the markers to the environment would be able to capture a large part of the response of the underlying QTL to stresses. $f(.)$ can then be rewritten as

$f(S,V)\xi = (S \otimes V_s)\varphi + g(S,V)\phi$, with φ effects of the factorial regression predictors and ϕ effects of the predictors from $g(S,V)$. As previously, a two-step approach was used. Estimation of β and α were performed using (Model 2) as for the two-step factorial regression. This restricted the machine learning task to the most complex part of the problem, the estimation of $g(S,V)$ from Model 9. Minimizing machine learning enables the use of classic predictors for β and α with well-known optimality and properties. To approximate $g(S,V)$, a non-linear function of unknown form with sparsity, I used soft rule fit, a modified version of the ensemble method RuleFit (Friedman and Popescu 2008; Akdemir and Heslot 2012). Soft rule fit has demonstrated good predictive ability for a number of machine learning tasks. It can capture non-linearity in the data as well as interactions between predictors (Friedman and Popescu 2008; Akdemir and Heslot 2012) in a sparse model without specifying a-priori the predictors for all the interactions or the shape of the response.

Ensemble learning is a relatively new approach to modeling, providing solutions to complex problems by combining simultaneously a number of models (Friedman and Popescu 2003). One of these methods, random forest (Breiman 2001) has already seen some applications in genetics (Bureau et al. 2005; Ogutu et al. 2011). Instead of identifying a single best performing model, the idea is to generate a very large number of predictors built on bootstrap samples of observations and variables. Those predictors are combined together using averaging (random forest) or penalized regression methods (Friedman and Popescu 2008) on the complete dataset.

This approach requires the definition of a family of models, used to generate predictors on bootstrap samples of observations and variables. Soft rule fit uses regression trees to generate predictors. Regression trees are a classic data-mining method that partitions the data into sets, each of which are simply modeled using

regression methods. The key aspect for the application here is that it groups observations based on the response variable and the predictors in a non linear way.

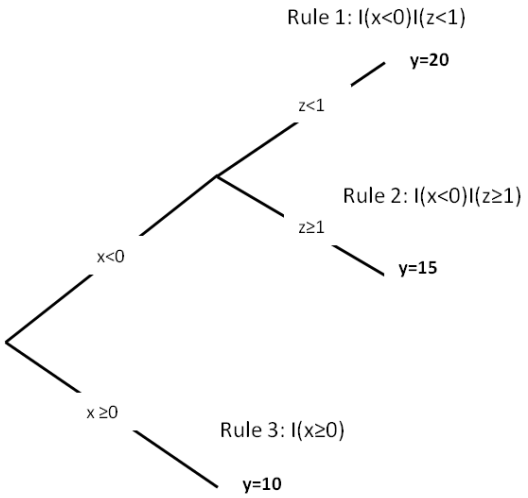


Figure 5.4. A simple regression tree built on a bootstrap sample of observations and variables. Each leaf node defines a rule which can be expressed as a product of indicator functions of half spaces. An indicator function $I(.)$ takes the value 1 if the condition it takes as input is true else it takes the value 0. Each rule specifies a 'simple' rectangular region in the input.

Figure 5.4 gives a graphical representation of a regression tree built on a bootstrap sample of observations and variables. For the response variable y and the predictors x and z, the regression tree algorithm identifies "splitting rules", defining nodes that partition the data. Figure 5.4 shows that the algorithm determined that the greatest variance reduction on the sample was obtained by dividing the data into two subsets based on whether x was positive or not. It was further identified that for x negative, the data was best modeled depending on whether z was superior or not to 1. This defined a complete partition of the observations based on response variable and predictors. This partition can be summarized by three binary rules that indicate to which group an observation belongs.

Soft rule fit uses those rules to derive a probabilistic assignment of each observation to a group using logistic regression. Each rule then provides a predictor taking values between 0 and 1 (Step 4 on Figure 5.3). The complexity of the rules is measured by the number of variables involved in a given rule. It is simply controlled by the number of observation groups allowed in the regression tree algorithm. This provides a simple way to control the level of complexity captured by the model as it fixes an upper bound on the number of variables involved in a given predictor.

By repeated sampling of the observations and variables, a matrix W of the soft rule predictors was obtained. W has dimension $mt \times h$ with h number of derived rules. W was further used as predictor of the residuals of the main effect model alone or with the factorial regression predictors described above and regressed on the residual of the main effect model using elastic net (Figure 5.3 step 5).

Model 9 can then be rewritten as follows:

$$y = 1_{mt}\mu + (I_t \otimes 1_m)\beta + (1_t \otimes V)\alpha + \theta\left[(S \otimes V_s)\varphi + W\phi\right] + \varepsilon_{10} \quad \text{(Model 10)}$$

With ϕ effects of the soft rule fit predictors. Model 10 can be fitted without the soft rule fit predictors (in which cases it reduces to Model 8) or without the factorial regression predictors W .

Model 10 was fitted in a two step procedure as described for models 6 and 7. Briefly $(S \otimes V_s)$ and W were simultaneously regressed with an elastic net on the residuals in model 2 to obtain φ and ϕ. θ, the scaling parameter, was then obtained by cross-validation on the training data. Higher prediction accuracy with the soft rules predictors included in the model (Model 10) compared to (Model 8) would provide evidence of non-linear effects.

About 5000 initial rules were derived from the data combining markers and stress covariates using the RuleFit algorithm (http://www-stat.stanford.edu/~jhf/r-

rulefit/rulefit3/R_RuleFit3.html). The maximum number of groups allowed in the regression trees was three or four. Rules were derived from bootstrap samples of the data using as response variable the residuals from Model 2. As a consequence stress covariates or markers alone might be associated with an apparent main effect on the bootstrap sample even if overall the data is corrected for the main environment and genotype effect. Of the rules generated, a large number combined only markers or only stress covariates. Such rules were removed to keep only rules combining one or several covariates with one or several markers, thus ensuring that both genotype and environment affected the rule.

Evaluation of model performance

To evaluate the performance of the models presented here, the accuracy was defined as the correlation between the predicted performance and the observed performance in a given environment. The main interest was in the capacity to discriminate between genotypes in unobserved environments. This is a simple way to assess the capacity of the model to capture G*E. To perform a valid statistical test, the data set was split randomly into two sets of 22 environments, balanced across years and locations with 4184 observations in the training set and 4840 observations in the validation set. There were 2195 genotypes in the validation set of which 544 were absent from the training set. The 22 validation environments were then predicted and the accuracy computed for each environment. The predictive ability of the model can then be assessed by the mean cross-validated accuracy across environments. Given this set-up, pairs of accuracies in the validation set are independent conditional on the training set and can be used to assess statistical significance of accuracy differences between models. For the pairs of correlations to be strictly (unconditionally) independent, separate sets of training environments would be required for each pair of correlations. Instead, each

pair of correlations derives from the same set of training environments. Therefore, the pairs of correlations are only independent *conditional on the chosen set of 22 training environments*. Thus, the statistical inference I can make is limited to the chosen set of 22 training environments. I cannot extend inferences to freshly chosen sets of training environments. This inference is quite limited and is justified by the data I show. What is tested is whether the model with weather data would do better than the baseline model on a new environment. A paired Wilcoxon rank sum test was used to assess the significance of the difference of the mean accuracies between models. The variance of the accuracy in the validation environments is an unbiased estimator of the variance of prediction accuracy and was reported as the coefficient of variation.

Inference about the genetic architecture of G*E

If the modeling of Q*E with stress covariates captures a significant part of the G*E variance and increases the model predictive power, it can be used to infer the genetic architecture of G*E. An importance measure can be derived for each rule and factorial regression predictor (Friedman and Popescu 2008) from model 10. All predictors were standardized to unit variance, and were continuous such that the importance measure was simply the absolute value of the coefficient of each G*E predictor (columns of $S \otimes V$ or W). From the importance of each of the predictors, the importance of the input variables (markers and stress covariates) can be derived as proposed by Friedman and Popescu (2008). It is computed as the sum of the importance of the predictors (soft rules and factorial regression) in which a given input variable appears, divided by the number of input variables involved in each predictor, such that, input variables involved in a predictor equally shared in its importance. The importance of the k^{th} stress covariate is then written:

$I_k = \sum_{i=1}^{s} \varphi_{101(k-1)+i} / 2 + \sum_{j=1}^{p} \phi_j / r_j$ with p numbers of rules including the k^{th} stress covariate and r_j the number of variables included in the j^{th} predictor including the k^{th} stress covariate. $\varphi_{101(k-1)+i}$ indicates the $101(k-1)+i$ elements of φ. A coefficient of one half is used for the factorial regression predictors because each of them has two input variables. Additionally, because each stress covariate was present in the model with three maturity-level parameterizations, importances were summed per stress covariate across those levels. The significance of the importance of the stress covariates was tested by permutations of the stress covariates between environments 100 times to generate a null distribution of the importance measure for each covariate. This permutation was done using the rules discovered on the non-permuted data to limit the computational load required. Using these rules produces a more stringent test than generating new rules using permuted data because some information from the real data is retained in these rules. Stress covariates were permuted in blocks between environments such that the covariance structure between stress covariates was preserved and each stress covariate had a single value in each environment. Then, soft rules and factorial regression predictors were generated using the permuted stress covariates and regressed on the residuals (Figure 5.3, step 5). A given stress covariate importance was considered significant if it was larger than the greatest importance obtained for that covariate with the permutations.

Model 10 allowed prediction of performance of any genotype in any environment based on the stress covariates and the markers Then, the G*E prediction term $(S \otimes V_s)\varphi + W\phi$ is an estimate of η, the environment-specific effect for each genotype from Model 4, even for unobserved environments. This estimate of η can be further used to estimate G the covariance matrix of genotype effects in environments from Model 3. For the best predictive model, the G*E term was predicted for all 2437

genotypes in all 44 environments. The derived table of predicted G*E response was used to estimate G. This corresponds to the predicted levels of genetic correlation between environments. Environments were clustered based on this covariance matrix using unweighted pair-group average agglomerative hierarchical clustering. Because of the machine learning predictors, there is no closed form estimate of G based on model coefficients as in random regression.

Results

Mixed model results

Using ASreml-R (Gilmour et al. 2009), Model 3 was fitted to generate simple variance component estimates. The additive genetic variance was estimated to be 6.5, the environment variance 172.2, the G*E variance 11.2 and the error variance 67.7. Heritability was computed using a generalized heritability measure suitable for unbalanced data using the predicted error variance from the mixed model (Piepho and Möhring, 2007) (formula 20) and was equal to 0.54.

Using ASReml-R, the size of the factorial regression problem (Model 5) quickly became intractable. It was not possible to fit genotype-specific sensitivity on the full dataset, even for one stress covariate, as the model required more than 250 gigabytes of RAM (Model 5) when covariance structures were included for u and γ. Fit was nevertheless possible when no covariance was included for u and γ. Focusing instead on marker sensitivities to the covariates seemed a suitable approach to decrease the dimensionality of the problem and it took advantage of powerful penalized regression methods (Model 6).

Marker variability

Focusing on marker sensitivity created a dimensionality problem. In the case of linear sensitivity, 129,987 additional predictors (1287 markers x 101 stress covariates) would have to be fitted in the model. To overcome that limitation, subsets of markers with particularly variable effects across environments were selected. The histogram plotted on a log scale in Figure 5.5 suggested that a few markers were extremely variable, the most variable being the marker for Ppd-D1, the main photoperiod sensitivity locus, with a variance of 0.043. The other markers with the largest variance were not

associated with any of the known major adaptation loci, despite the inclusion of diagnostic markers for those loci in the analysis.

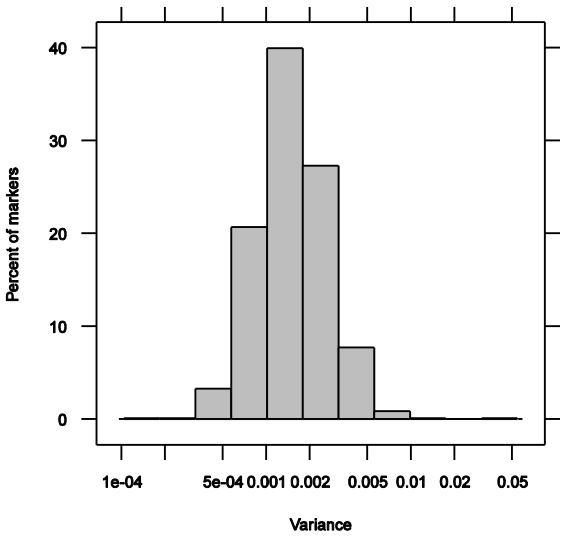


Figure 5.5. Distribution of the variance of marker effects, computed in each environment, across environments, plotted on a log scale.

Two-step approach results

Despite an appealing simplicity, using the SGL to perform a one-step analysis was too computationally intensive to be feasible. Thus, all the results presented here are from the two-step approach (Model 10).

Figure 5.6 presents the mean prediction accuracies for different models capturing G*E compared to the base model with no modeling of G*E included in cross validation. Results indicated a gain in mean prediction accuracy (0.25 to 0.277) compared to the base model accuracy when predictors of G*E were included. Most of the gain in accuracy came from the linear response of markers to the stress covariates. With all three models considered, accuracy reached a maximum when 250 markers were included for factorial regression.

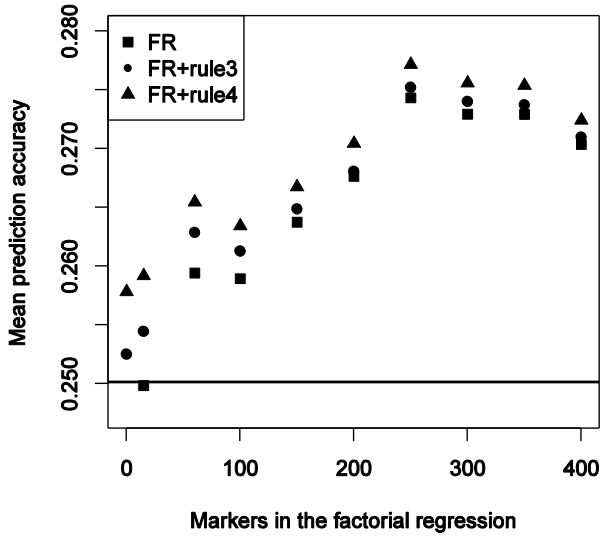


Figure 5.6. Predictive performances of the models as a function of the number of markers for which a linear sensitivity to each covariate is fitted from the cross-validation. Three scenarios are plotted, with no inclusion (square) or inclusion of soft rule fit predictors of order three (circle) or order four (triangle). Predictive performance is measured as the mean prediction accuracy for the 22 environments removed from the training set. The first points with 0 abscissa correspond to a model with rules only.

The best model in cross-validation comprised 1584 soft rules, each containing at least one stress covariate. The best model (with soft rules of order four and 250 factorial regression predictors) provided an 11.1% increase in accuracy and a 10.8% decrease in accuracy coefficient of variation over Model 2. On the same cross-validation settings, the G*E model captured on average 3.7% of the variance of residuals of the base model in each validation environment. Inclusion of the soft rule fit predictor improved accuracy slightly, for any number of markers included. The best model (with soft rules of order four and 250 factorial regression predictors) was significantly better than a model with 250 factorial regression predictors and no rules included (P-value = 0.093). This indicated that part of the G*E response was due to a non-linear

response of the QTL to the environment. When stress covariates were permuted between environments, including a G*E term decreased the predictive ability of the model. This result rules out overfitting by the model.

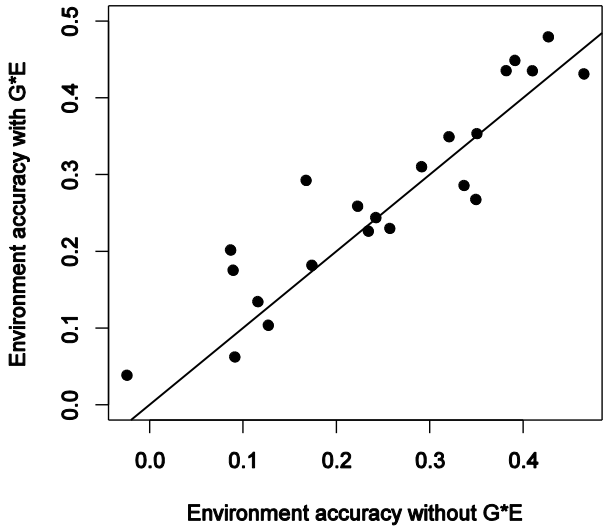


Figure 5.7. Comparison of the prediction accuracy in each predicted environment from the cross validation between the main effect model and the model with rules of order four and 250 markers for the factorial regression added to the main effect prediction. The line indicates the identity.

Figure 5.7 presents the detailed cross-validation results for the model with rules of order four and 250 markers in the factorial regression. The line corresponds to the identity such that if a validation environment is over the line, there is a prediction gain by modeling G*E for that environment. Residuals of a few environments were poorly predicted by the model, as expected under the Anna Karenina effect.

Inference of the genetic architecture of G*E

All the following results are based on the model with rules of order four and 250 markers for the factorial regression, which was the most predictive model in cross-validation.

Six hundred eighty nine markers (53.5% of all markers) had an importance different from 0 and thus were included in the model to predict G*E. However, inclusion does not mean significance.

The most important marker was the marker for *ppd-D1*, the photoperiod sensitivity locus with an importance twice the importance of the second most important marker. Ninety-nine percent of this importance was due to the factorial regression predictors and not to the rules, indicating that the effect of the *ppd-D1* locus on yield changes linearly with stress. This marker was also the most variable marker across environments. However apart from *ppd-D1*, there was no correlation between variability across environment and importance based on the factorial regression. This could be explained by stresses not taken into account by the covariates, or this can point out the inefficiency of the marker selection procedure. The *ppd-D1*photoperiod insensitive allele had a mean frequency of 51.8% in the different environments with a minimum of 8.6% and a maximum of 76.8%. When *ppd-D1* was fitted alone in a factorial regression model, the cross-validated accuracy was equal to the accuracy obtained when fitting Model 1 alone.

Other perfect markers for vernalization, photoperiod sensitivity, or dwarf status had little or no importance. The other most important markers did not correspond to known loci affecting phenology. There was no correlation between the marker importance and their main effect on yield. Similarly the marker importance was not correlated to their main effect on heading date (data not shown).

All stress covariates were included in the soft rule terms. This is evidence of the complexity of the G*E response, as it involves all the stress covariates in the model covering abiotic stresses over the whole plant cycle. Despite this overall complexity, only ten stress covariates had an importance larger than the importance observed in 100 permutations and are presented in Table 5.2.

Table 5.2. Importance of the eight stress covariates with a significant importance, for the model with rules of order 4 and 250 markers for the factorial regression. Importances are rescaled to give a score of 100 to the largest one.

Stress covariate	stress type	Importance
stmpmf	Sum temperature meiosis to flowering	100.00
ntddr	Drought early spring	88.97
spetpe1	Drought early spring	84.99
st25ef	Heat stress before flowering	79.69
nd25ef	Heat stress before flowering	77.53
st25fh	Heat stress early grain filling	76.25
nd25fh	Heat stress early grain filling	71.65
latitude	North/South trend	29.52

From those most important stress covariates, a clear picture emerges about the stress creating the most G*E on winter wheat in France. Almost all the stress covariates presented in Table 5.2 related to stresses before flowering such as drought stress in early spring (ntddr, spetpe1) and heat stress before flowering (stmpmf, st25ef, nd25ef). Heat stress at the beginning of grain filling was also important (st25fh, nd25fh). Latitude captured a North/South gradient in weather patterns. This is evidence that the most critical stage for G*E and abiotic stress sensitivity unexplained by geography are stresses before flowering and not the late stage stresses.

As the model enabled the prediction of part of the G*E response in each environment based on markers and stress covariates, fitted values were calculated for all genotypes all environments in the dataset. However, this could also be done for any environment with daily weather data. Those predicted values could be used to study the stability of genotypes in a set of environments. Consequently, for each of the 2437 genotypes, the variance of the predicted G*E response in a set of environments was compared to the main genotype effects and there was no correlation. Those fitted values were also used

to compute a predicted G*E correlation matrix between environments. This correlation matrix corresponds to the genetic correlation between environments as captured by the model (Figure 5.8).

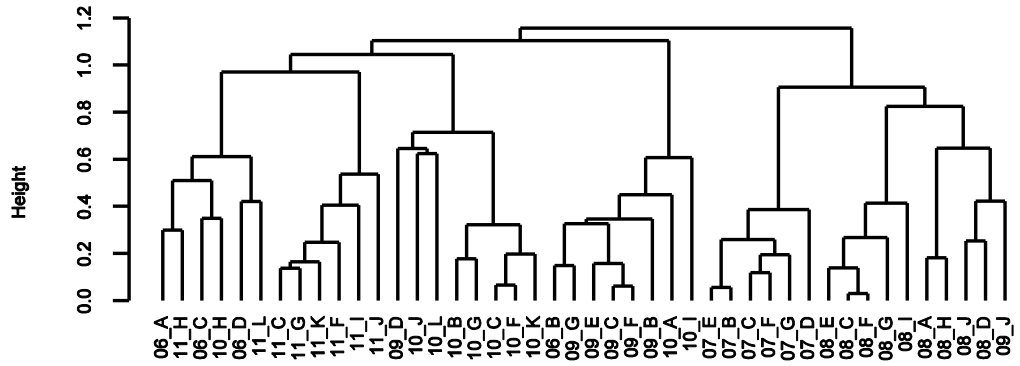


Figure 5.8. Hierarchical agglomerative clustering of the environments based on the predicted G*E response of all genotypes. Environments were named using the last two digits of the year followed by a location code.

The correlation matrix indicated divergent G*E response between environments with a wide range of correlations between environments. Several clusters of environments corresponding mostly to year, were identifiable on the dendrogram (Figure 5.8). Those clusters were also confirmed by looking at the heatmap of the correlation matrix (Figure 5.9). They corresponded mostly to a clustering by year. Clear clusters can be identified for 2007, 2008 and 2011. The cluster of environments on the left of the dendrogram spanned several years and corresponded mostly to locations in the south of France. A similar cluster analysis was performed directly using the stress covariates produced a pattern that was less differentiated by year and some outlier environments that disappeared in the clustering approach based on predicted G*E (data not shown).

Discussion

A disappointing gain in accuracy?

My model was able to predict part of the G*E response of genotypes in unobserved environments. Though the gain was statistically significant, it was small. Figure 5.7, however, indicates that gains were much larger in some environments, especially environments where prediction accuracy was low with the baseline model. In a survey of the G*E literature I found only two papers (Burgueno et al. 2011; Burgueno et al. 2012) concerned with predicting G*E to report cross-validation results. Papers I identified using factorial regression such as Crossa et al. 1999 usually only reported a model fit to the whole dataset and did not assess prediction accuracy.

The average gain I observed (11.1%) was higher than the one reported in Burgueno et al. (2011). They reported a gain of 6% on average across 6 datasets when using a factor analytic model. Analysis of results presented in Burgueno et al. (2012) suggested an average accuracy gain of 7.2% (0.44 to 0.471) when predicting genotypes absent from the training set and a gain of 19.6% (0.474 to 0.55) when predicting genotypes present in the training set in a different environment using a factor analytic model and the realized relationship matrix. For their model to be predictive some genotypes needed to be observed in the environment to be predicted.

Here I am predicting G*E deviation for unobserved environments which is clearly a more difficult task. The cross-validation setting is rather stringent here with the training dataset effectively reduced to half the size of the total dataset. With more data, in particular more environments, I would expect better performance. I also show that the gain in accuracy is statistically significant and that when using permuted covariates the model has no predictive power. These results together indicate that the model is picking up real G*E signal in the data.

It is also important to remember that the reported accuracies are correlations between predicted values and phenotypes measured in a single environment with a low number of replicates and consequently a low repeatability. Based on the variance components estimated from Model 3, the within environment repeatability should be approximately $\left(\sigma_u^2 + \sigma_{ge}^2\right) \Big/ \left(\sigma_u^2 + \sigma_{ge}^2 + \sigma_e^2\right) = 0.207$ with σ_u^2, the additive genetic variance, σ_{ge}^2 the G*E variance and σ_e^2 the error variance estimated with Model 3. So the maximum accuracy would be about 0.455 if all genetic and G*E variance was predicted.

Strategies integrating statistical and crop growth models for phenotype prediction

Levins (1966) states that in building models there is a tension between the goals of realism, generality, and accuracy making it impossible to create a model that fulfills all goals simultaneously. Purely statistical models (e.g., regression and machine learning) sacrifice realism and generality in favor of accuracy. These models minimize prediction error but usually cannot be extrapolated to conditions outside those previously observed, and their parameters have little interpretive value relative to the underlying biology of the problem. They are pure black box models. Crop growth models, in contrast, sacrifice accuracy in favor of a certain level of realism and generality. These models do not completely lack predictive power. They can indicate which conditions will increase or decrease performance and serve best to provide qualitative rather than quantitative, error-minimizing, predictions. For the problem of G*E prediction in breeding, I require accuracy because I will seek to select on the basis of model predictions. Generality is also required because the environments that interest us are those of the future and are therefore, by definition, unobserved. Finally, while model realism is probably relegated to the lowest priority, I do not want to

discard it entirely because interpretation of the model can guide future experimental and selection efforts.

In broad terms, I place crop growth and statistical models in series with the outputs of the former providing the inputs of the latter. Thus, I aim for the crop growth model to provide some level of generality and realism, first by condensing massive quantities of weather variables in a limited number of covariates and second by tying those covariates to phenologically relevant and interpretable plant stresses. The statistical model must then provide predictive accuracy. I note that placing the models in the opposite order (statistical then crop growth) is also a possibility. In that case, the statistical model would predict crop growth model parameters and the crop growth model would then combine those with weather data to produce a prediction. For that strategy, the crop growth model parameters are used as traits (Reymond et al. 2004; Reymond et al. 2003), which would require specific phenotyping experiments. The strategy that I use requires only phenotypic data generated by a breeding program for the purpose of selection. In addition, as discussed above, there are concerns that crop models are not sensitive enough to capture the subtle performance differences between elite genotypes (White et al. 2008).

In this paper I integrated environment data in the analysis using a crop model as a tool to generate metadata about the trial that included phenology. This is a key point because phenology data is not usually collected in plant breeding trials, with the possible exception of heading date. Furthermore, the determination of most of the developmental stages is difficult and labor intensive. Once development stages are known, agronomy and plant physiology knowledge can be leveraged to define stress covariates by stage. This has multiple advantages because it reduces the dimensionality of the data to a few dozen covariates. It also enables use of the large datasets generated by plant breeding activities to study G*E. While a major data

reduction was achieved, there were still more predictors (environmental covariates) than observations (environments). Here, only weather data was used but the framework developed could accommodate other kinds of environment variables such as soil quality types and disease pressure. This type of data reduction strategy with further use in QTL mapping was first proposed by Boer et al. (2007). Here, I extended it to a large number of stress covariates and genome-wide markers while capturing non-linearity of responses.

Inference about the genetic architecture of G*E

The best performing model included predictors for linear responses to the stress covariates as well as soft rule fit predictors capturing non-linearity. This suggests that part of the G*E is not amenable to modeling by a linear response. In addition, the use of a crop model and the definition of stress covariates by growth stage also captured some non-linearity. For example heat stress is expected to be critical at flowering and less so earlier in the cycle. This creates major non-linearity issues in using weather data for modeling directly. The use of stress covariates provides a simplification of the problem which is difficult to quantify.

Assuming that the model captured enough G*E variance to provide useful insight in the genetic architecture of G*E, the use of biologically meaningful stress covariates facilitated the interpretation of the model. The significantly important stress covariates were related to radiation and water stress before flowering rather than to terminal stresses. This has important implications because it suggests that breeding efforts for stress tolerance should focus more on those specific stresses. Alternatively, as terminal stress was expected a-priori to be important, it could suggest that the breeding program from 2006 to 2011 did not sample environments with terminal stress. Despite

an expected large north-south G*E pattern, results indicated that annual rather than latitudinal variation was more important.

One of the most important loci for G*E was *Ppd-D1,* a photoperiod sensitivity locus, indicating that a major determinant of the G*E response in this dataset is phenology. However, *Ppd-D1* alone did not capture a significant part of the G*E variance. Other important loci for G*E did not correspond to any of the other known major adaptation genes and these loci warrant further investigation. Results suggest that Q*E is pervasive and characterized by small interactions among large numbers of regions of the genome and a large number of stresses as expected under the Anna Karenina effect. The importance of the markers for G*E prediction was not related to their main effect and the genotype main effect was not related to the variability of G*E response. These results also suggest that genotypes can be selected for stability without penalizing performance. However, the model does not explain a large part of the G*E variance. If the alternative hypothesis of a correlation between main and interaction effects holds, I do not know what power I would have to detect it. Stability was defined here as the variance of the predicted G*E for each genotype. Other definitions are possible. Stability is not necessarily a desirable trait if it means consistently poor performance. Capacity of genotypes to take advantage of good growing conditions is a favorable trait which is potentially associated with performance instability according to the definition I used.

My results indicated that the QTLs causing the most Q*E have small main effects. This suggests that focusing on markers with large overall main effects when trying to identify Q*E is inefficient.

I hypothesize that those QTLs with a large Q*E effect on yield are not likely to be detected in mapping experiments across environments because their measured main effect will not be consistent across environments. If they are detected within-

environment, they are unlikely to be validated in separate mapping experiments in different environments. If identifying consistent QTL main effects is the goal, those QTLs might only be detected when focusing directly on an underlying physiology or plant architecture trait. For example, in this dataset ppd-D1 did not have a consistent main effect on yield across environments, but would have been detected if the mapping focused on the main effect of photoperiod sensitivity.

A new tool to deal with G*E in breeding programs

Historical weather data or predicted weather data from climate change models could be used in simulations to investigate the target population of environments (TPE). The TPE is the mixture of environments expected for the intended region of production (Comstock 1977). In most cases, the composition of the TPE is unknown. Simulation studies show that in the case of cross-over G*E it is beneficial to weight the trials by their expected frequency of occurrence in the TPE (Podlich et al. 1999). My approach provided an accessible way to determine those frequencies. By predicting genotype performance using historical weather records, the frequency of occurrence of the clusters identified in Figures 5.8 and 5.9 can be calculated. This could be used to optimize the phenotypic testing strategy. Most of the environment clustering I observed was by year suggesting that my sample of environments, while large, was not large enough to cover all expected environment types. Using this data only provided a partial glimpse of the TPE. This interpretation supposes that the locations are a spatially representative sample of the TPE.

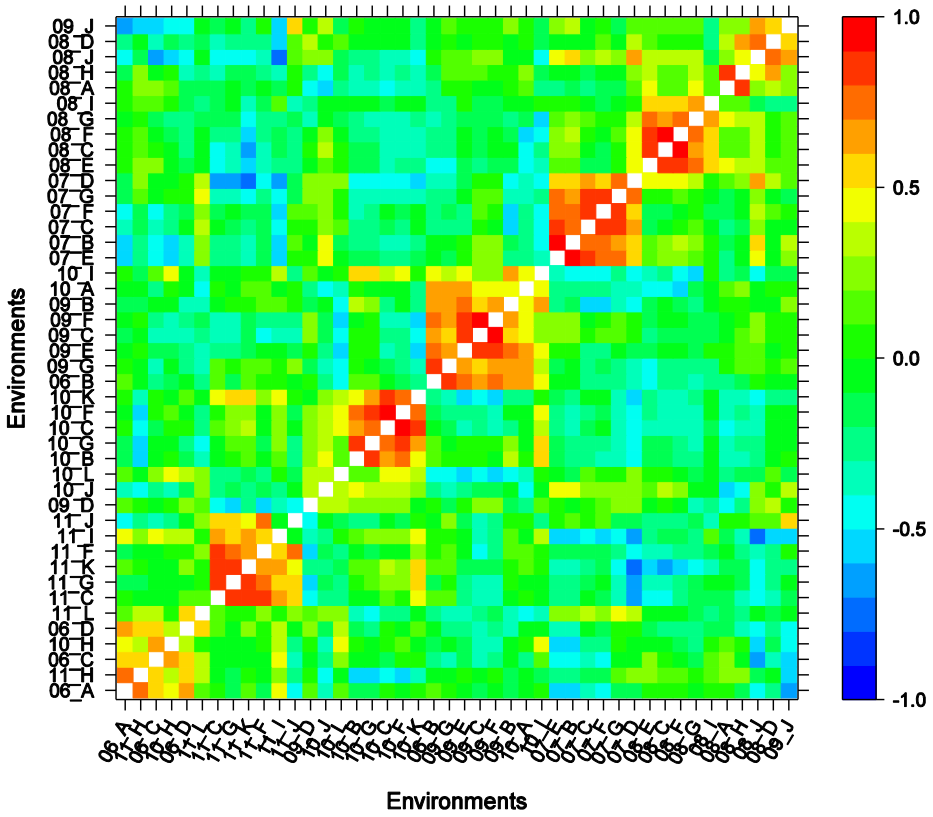


Figure 5.9. Heatmap of the correlation of G*E predicted values between environments after hierarchical agglomerative clustering of the environments based on the predicted G*E response of all genotypes. Environments were named using the last two digits of the year folloI d by a location code.

Using the predicted G*E response instead of the stress covariates to cluster environments allowed clustering on the predicted level of genetic correlation between environments, which is the parameter of interest for breeding purposes. This was possible even for environments with no phenotypic data. Multiple stresses were considered and none of them was suspected to be the main cause of G*E. Consequently, it was not meaningful to directly use the stress covariates to group environments as in previous studies (Chapman et al. 2000a; Chapman et al. 2000b; Chapman et al. 2000c). In their studies, the main cause of G*E was drought stress, such that environments could be clustered based on the pattern of drought stress.

Using the predicted G*E response also captured non-linear responses of genotypes to stresses and threshold effects.

By leveraging agronomic knowledge and the large historical datasets generated by breeding programs, this new model provides insight into the genetic architecture of genotype by environment interactions and predicts genotype performance based on past and future weather scenarios. The model can therefore provide a better knowledge of the current and future TPE. This knowledge should translate to an improved design of phenotypic testing strategies.

Acknowledgments.

I thank Pierre Martre for providing the crop model. The reviewers provided excellent comments that significantly improved the paper. JRC-MARS - Meteorological Data Base - EC – JRC provided access to the interpolated meteorological data. This research was supported in part by USDA-NIFA-AFRI grants, award numbers 2009-65300-05661, 2011-68002-30029, and 2005-05130 and by Hatch project 149-449. Limagrain Europe provided financial support for N. Heslot.

References

Akdemir D, Heslot N (2012) Soft rule ensembles for statistical learning. Arxiv preprint arXiv: 1205.4476

Boer MP, Wright D, Feng L, et al. (2007) A mixed-model quantitative trait loci (QTL) analysis for multiple-environment trial data using environmental covariables for QTL-by-environment interactions, with an example in maize. Genetics doi: 10.1534/genetics.107.071068

Brancourt-Hulmel M, Denis JB, Lecomte C (2000) Determining environmental covariates which explain genotype environment interaction in winter wheat through probe genotypes and biadditive factorial regression. Theor Appl Genet doi: 10.1007/s001220050038

Brancourt-Hulmel M, Lecomte C, Meynard JM (1999) A diagnosis of yield-limiting factors on probe genotypes for characterizing environments in winter wheat trials. Crop Sci doi: 10.2135/cropsci1999.3961798x

Breiman L (2001) Random forests. Mach Learn doi: 10.1023/A:1010933404324

Breiman L, Friedman J (1985) Estimating optimal transformations for multiple regression and correlation. J Am Statist Assoc 80:580–598.

Bureau A, Dupuis J, Falls K, et al. (2005) Identifying SNPs predictive of phenotype using random forests. Genet epidemiol doi: 10.1002/gepi.20041

Burgueño J, Crossa J, Cornelius PL, Yang RC (2008) Using factor analytic models for joining environments and genotypes without crossover genotype × environment interaction. Crop Sci doi: 10.2135/cropsci2007.11.0632

Burgueño J, Crossa J, Cotes JM, et al. (2011) Prediction Assessment of Linear Mixed Models for Multienvironment Trials. Crop Sci doi: 10.2135/cropsci2010.07.0403

Burgueño J, De los Campos G, Weigel K, Crossa J (2012) Genomic Prediction of breeding values when modeling genotype × environment interaction using pedigree and dense molecular markers. Crop Sci doi: 10.2135/cropsci2011.06.0299

Chapman SC, Cooper M, Butler D, Henzell R (2000a) Genotype by environment interactions affecting grain sorghum. I. Characteristics that confound interpretation of hybrid yield. Aust J Agr Res doi: 10.1071/AR99020

Chapman SC, Cooper M, Hammer G, Butler D (2000b) Genotype by environment interactions affecting grain sorghum. II. Frequencies of different seasonal patterns of drought stress are related to location effects on hybrid yields. Aust J Agr Res 51:209–221.

Chapman SC, Hammer G, Butler D, Cooper M (2000c) Genotype by environment interactions affecting grain sorghum. III. Temporal sequences and spatial patterns in the target population of environments. Aust J Agr Res doi: 10.1071/AR99022

Chenu K, Chapman SC, Hammer G, et al. (2008) Short-term responses of leaf growth rate to water deficit scale up to whole-plant and crop levels: an integrated modelling approach in maize. Plant, Cell Environ. doi: 10.1111/j.1365-3040.2007.01772.x

Chenu K, Deihimfard R, Chapman SC (2013) Large-scale characterization of drought pattern: a continent-wide modelling approach applied to the Australian wheatbelt - spatial and temporal trends. The New phytol. doi: 10.1111/nph.12192

Chiquet J, Grandvalet Y, Charbonnier C (2012) Sparsity with sign-coherent groups of variables via the cooperative-Lasso. Ann. Appl. Stat doi: 10.1214/11-AOAS520

Comstock RE (1977) Quantitative genetics and the design of breeding programs. In: Pollak E, Kempthorne O, Bailey TB (eds) Proceedings of the International Conference on Quantitative Genetics. Iowa State University Press, Ames IA, pp 705–718

Cooper M, DeLacy IH (1994) Relationships among analytical methods used to study genotypic variation and genotype-by-environment interaction in plant breeding multi-environment experiments. Theor Appl Genet doi: 10.1007/BF01240919

Crossa J, Vargas M, Van Eeuwijk FA, et al. (1999) Interpreting genotype × environment interaction in tropical maize using linked molecular markers and environmental covariables. Theor Appl Genet doi: 10.1007/s001220051276

Cullis BR, Smith a. B, Beeck CP, Cowling W a (2010) Analysis of yield and oil from a series of canola breeding trials. Part II. Exploring variety by environment interaction using factor analysis. Genome doi: 10.1139/G10-080

DeLacy IH, Basford KE, Cooper M, et al. (1996) Analysis of multi-environment trials - An historical perspective. In: Cooper M, Hammer G (eds) Plant adaptation and crop improvement. CAB International, Wallingford, UK, pp 39–124

Demotes-Mainard S, Doussinault G, Meynard JM (1996) Abnormalities in the male developmental programme of winter wheat induced by climatic stress at meiosis. Agronomie doi: 10.1051/agro:19960804

Denis JB (1988) two-way analysis using covariates. Statistics 19:123–132.

Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics, 4th ed. Pearson, Prentice Hall, Harlow, UK

Fischer RA (1985) Number of kernels in wheat crops and the influence of solar radiation and temperature. J Agri Sci doi: 10.1017/S0021859600056495

Friedman JH, Hastie T (2010) A note on the group lasso and a sparse group lasso. Arxiv preprint arXiv:10010736

Friedman JH, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. J Stat Softw 33:1.

Friedman J, Popescu BE (2003) Importance sampled learning ensembles. J Mach Learn Res 94305:1–32.

Friedman JH, Popescu BE (2008) Predictive learning via rule ensembles. Ann. Appl. Stat. 2:916–954.

Gallagher JN, Biscoe PV (1978) Radiation absorption, growth and yield of cereals. J Agri Sci doi: 10.1017/S0021859600056616

Gate P (1995) Ecophysiologie du blé. De la plante à la culture. Tec & Doc, Paris, France. pp 430.

Gauch HG (2006) Statistical analysis of yield trials by AMMI and GGE. Crop Sci doi: 10.2135/cropsci2005.07-0193

Gianola D, Fernando RL, Stella A (2006) Genomic-assisted prediction of genetic value with semiparametric procedures. doi: 10.1534/genetics.105.049510

Gilmour AR, Gogel B, Cullis BR, et al. (2009) ASREML user guide release 3.0. VSN International Ltd, Hemel Hempstead, UK

Habier D, Fernando RL, Dekkers JCM (2007) The impact of genetic relationship information on genome-assisted breeding values. Genetics doi: 10.1534/genetics.107.081190

Hammer G, Kropff MJ, Sinclair TR, Porter JR (2002) Future contributions of crop modelling—from heuristics and supporting decision making to understanding genetic regulation and aiding crop improvement. Eur J Agron doi: 10.1016/S1161-0301(02)00093-X

He J, Le Gouis J, Stratonovitch P, et al. (2012) Simulation of environmental and genotypic variations of final leaf number and anthesis date for wheat. Eur J Agron doi: 10.1016/j.eja.2011.11.002

Heffner EL, Lorenz A. J, Jannink J-L, Sorrells ME (2010) Plant breeding with genomic selection: gain per unit time and cost. Crop Sci doi: 10.2135/cropsci2009.11.0662

Heslot N, Jannink J-L, Sorrells ME (2013) Using genomic prediction to characterize environments and optimize prediction accuracy in applied breeding data. Crop Sci doi: 10.2135/cropsci2012.07.0420

Hunt LA (1991) Postanthesis temperature effects on duration and rate of grain filling in some winter and spring wheats. Can J Plant 617:609–617.

Jamieson PD, Semenov MA, Brooking IR, Francis GS (1998) Sirius: a mechanistic model of wheat response to environmental variation. Eur J Agron doi: 10.1016/S1161-0301(98)00020-3

Jullien A, Mathieu A, Allirand JM, et al. (2011) Characterization of the interactions between architecture and source-sink relationships in winter oilseed rape (Brassica napus) using the GreenLab model. Ann bot-London doi: 10.1093/aob/mcq205

Kelly AM, Cullis BR, Gilmour AR, et al. (2009) Estimation in a multiplicative mixed model involving a genetic relationship matrix. Genet Sel Evol doi: 10.1186/1297-9686-41-33

Landau S, Mitchell RA, Barnett V, et al. (2000) A parsimonious, multiple-regression model of wheat yield response to environment. Agr Forest Meteorol doi: 10.1016/S0168-1923(99)00166-5

Landau S, Mitchell RA, Barnett V, et al. (1998) Testing winter wheat simulation models' predictions against observed UK grain yields. Agr Forest Meteorol doi: 10.1016/S0168-1923(97)00069-5

Lecomte C (2005) Experimental evaluation of varietal innovations. Proposition of genotype - environment analysis tools adapted to the diversity of needs and constraints of the professionals of the seeds industry. Dissertation, AgroParisTech. pp 262.

Levins R (1966) The strategy of model building in population biology. Am Sci 54:421–431.

Löffler CM, Wei J, Fast T, et al. (2005) Classification of maize environments using crop simulation and geographic information systems. Crop Sci doi: 10.2135/cropsci2004.0370

Lorenz AJ, Chao S, Asoro FG, et al. (2011) Genomic selection in plant breeding: knowledge and prospects. Adv Agron doi: 10.1016/B978-0-12-385531-2.00002-5

Ma CX, Casella G, Wu R (2002) Functional mapping of quantitative trait loci underlying the character process: a theoretical framework. Genetics 161:1751–62.

Malosetti M, Voltas J, Romagosa I, et al. (2004) Mixed models including environmental covariables for studying QTL by environment interaction. Euphytica doi: 10.1023/B:EUPH.0000040511.46388.ef

Martre P, Jamieson PD, Semenov MA, et al. (2006) Modelling protein content and composition in relation to crop nitrogen dynamics for wheat. Eur J Agron doi: 10.1016/j.eja.2006.04.007

Messina C, Hammer G, Dong Z, et al. (2009) Modelling crop improvement in a GXEXM framework via gene-trail-phenotype relationships. In: Sadras VO, Calderini D (eds) Crop physiology: applications for genetic improvement and agronomy, Elsevier. Netherlands, pp 235–265

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829.

Meynard JM, Sebillotte M (1994) L'élaboration du rendement du blé, base pour l'étude des autres céréales à paille. In: Picard D, Combe L (eds) Elaboration du rendement des principales cultures annuelles. INRA, Paris, pp 31–51

Monteith J (1972) Solar radiation and productivity in tropical ecosystems. J. Appl. Ecol 9:747–766.

Ogutu JO, Piepho HP, Schulz-Streeck T (2011) A comparison of random forests, boosting and support vector machines for genomic selection. BMC Proc doi: 10.1186/1753-6561-5-S3-S11

Park T, Casella G (2008) The bayesian lasso. Amer. Statist. Assoc. doi: 10.1198/016214508000000337

Pérez P, De los Campos G, Crossa J, Gianola D (2010) Genomic-enabled prediction based on molecular markers and pedigree using the bayesian linear regression package in R. Plant Gen doi: 10.3835/plantgenome2010.04.0005

Piepho HP (1998) Empirical best linear unbiased prediction in cultivar trials using factor-analytic variance-covariance structures. Theor Appl Genet doi: 10.1007/s001220050885

Piepho HP, Denis JB, Van Eeuwijk FA (1998) Predicting cultivar differences using covariates. J Agric Biol Envir S doi: 10.2307/1400648

Piepho HP, Möhring J (2006) Selection in Cultivar Trials—Is It Ignorable? Crop Sci doi: 10.2135/cropsci2005.04-0038

Piepho, HP, Möhring, J. (2007). Computing heritability and selection response from unbalanced plant breeding trials. Genetics doi:10.1534/genetics.107.074229

Piepho HP, Möhring J, Melchinger AE, Büchse A (2008) BLUP for phenotypic selection in plant breeding and variety testing. Euphytica doi: 10.1007/s10681-007-9449-8

Piepho HP, Ogutu JO, Schulz-Streeck T, et al. (2012) Efficient computation of ridge-regression best linear unbiased prediction in genomic selection in plant breeding. Crop Sci doi: 10.2135/cropsci2011.11.0592

Podlich DW, Cooper M, Basford KE (1999) Computer simulation of a selection strategy to accommodate genotype-environment interactions in a wheat recurrent selection programme. Plant Breeding doi: 10.1046/j.1439-0523.1999.118001017.x

Quilot B, Génard M, Kervella J, Lescourret F (2004) Analysis of genotypic variation in fruit flesh total sugar content via an ecophysiological model applied to peach. Theor Appl Genet doi: 10.1007/s00122-004-1651-7

Reymond M, Muller B, Leonardi A, et al. (2003) Combining quantitative trait loci analysis and an ecophysiological model to analyze the genetic variability of the responses of maize leaf growth to temperature and water deficit. Plant Physiol doi: 10.1104/pp.013839.soil

Reymond M, Muller B, Tardieu F (2004) Dealing with the genotype x environment interaction via a modelling approach: a comparison of QTLs of maize leaf length or width with QTLs of model parameters. J Exp Bot doi: 10.1093/jxb/erh200

Smith AB, Cullis BR, Thompson R (2005) The analysis of crop cultivar breeding and evaluation trials: an overview of current mixed model approaches. J Agr Sci doi: 10.1017/S0021859605005587

Sofield I, Evans L, Cook M, Wardlaw I (1977) Factors influencing the rate and duration of grain filling in wheat. Aust J Plant Physiol doi: 10.1071/PP9770785

Stone P, Nicolas M (1998) The effect of duration of heat stress during grain filling on two wheat varieties differing in heat tolerance: grain growth and fractional protein accumulation. Aust J Plant Physiol doi: 10.1071/PP96114

Tashiro T, Wardlaw I (1990) The response to high temperature shock and humidity changes prior to and during the early stages of grain development in wheat. Aust J Plant Physiol doi: 10.1071/PP9900551

Van der Goot E, Orlandi S (2003) Technical description of interpolation and processing of meteorological data in CGMS. Joint Research Centre of the European Commission, Ispra, Italy, pp 23.

Van Eeuwijk FA, Denis J-B, Kang MS (1996) Incorporating additional information on genotypes and environments in models for two-way genotype by environments tables. In: Kang MS, Gauch HG (eds) Genotype-by-environment interaction. CRC Press, Boca Raton,FL, pp 15–50

Van Eeuwijk FA, Malosetti M, Yin X, et al. (2005) Statistical models for genotype by environment data: from conventional ANOVA models to eco-physiological QTL models. Aust J Agr Res doi: 10.1071/AR05153

White JW, Herndl M, Hunt LA, et al. (2008) Simulation-based analysis of effects of loci on flowering in wheat. Crop Sci doi: 10.2135/cropsci2007.06.0318

Windhausen VS, Wagener S, Magorokosho C, et al. (2012) Strategies to subdivide a target population of environments: results from the CIMMYT-led maize hybrid testing programs in Africa. Crop Sci doi: 10.2135/cropsci2012.02.0125

Zadoks JC, Chang TT, Konzak CF (1974) A decimal code for the growth stages of cereals. Weed Res doi: 10.1111/j.1365-3180.1974.tb01084.x